Influence of Genetic Variance on an Occupational Exposure Assessment Model of 1,6-Hexamethylene Diisocyanate

Kathie Sun

A Master's thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Public Health in the Department of Environmental Sciences and Engineering in the Gillings School of Global Public Health.

Chapel Hill 2016

Approved by,

Advisor: Leena A. Nylander-French

Reader: John E. French

Reader: Samir Kelada

Reader: Joachim Pleil

Table of Contents

Abstract

Significant differences in systemic response to xenobiotic exposure result from inter-individual genetic variation, but this variation is not included as a predictor of outcome in exposure assessment models. We developed an approach to investigate and identify individual differences in genetic variation that influence biomarkers of exposure levels. 1,6-Hexamethylene diamine (HDA) was measured in collected samples of blood and urine as a quantitative biological phenotype in a well-characterized population of 33 automotive spray painters exposed to 1,6-hexamethylene diisocyanate (HDI). Our statistical modeling approach contains whole-genome markers along with exposure predictors to determine the contribution of individual genetic variants and their interactions to the observed biomarker levels among the exposed workers. A total of 25 single nucleotide polymorphisms were significantly associated with measured HDA biomarker levels in urine and blood after controlling for multiple comparisons at a false discovery rate $q<0.20$. The genetic marker most associated with urine biomarker levels, rs169, was also a significant predictor in linear mixed-effects models that accounted for personal HDI exposure across multiple visits per worker ($p<0.05$). Our results indicate that the incorporation of genetic markers informs exposure assessment models for HDI.

Acknowledgements

I would like to express my heartfelt appreciation to everyone that contributed to the completion of my Master's thesis. My advisor, Leena Nylander-French, was always willing to help and to provide guidance for my project. Above all, Leena was encouraging and optimistic which never failed to inspire me to stay focused. I owe so much of my project to data collected by previous members of the Nylander-French lab, including Kenneth Fent, Sheila Flack, Linda Gaines, and Jennifer Thomasen. Thank you all for completing the hard work so I didn't have to! I could not have finished this project without the help of Rong Jiang, whose methods I adapted and who guided me through the dangerous terrain of statistical analysis. In addition, I sought help on questions about statistics and linear regression from every biostatistician that I came across: a thank you to my professors, teaching assistants, even my PhD-candidate roommate for putting up with my endless questions. My committee members also contributed valuable advice: Jef French for running the GeneMANIA analysis, Samir Kelada for providing insightful feedback on analyzing genetic associations, and Joachim Pleil for sharing his knowledge of exposure science.

This study was supported by grants from the National Institute for Occupational Safety and Health (R21-OH010203, R01-OH007598, T42/CCT422952, and T42/OH-008673).

Finally, I would like to give a big shout out to the wonderful community of environmental scientists at the UNC School of Public Health for supporting me and each other in our academic endeavors. The future is bright with so many dedicated scholars working on the imminent problems of our time.

List of Tables

List of Figures

List of Abbreviations

ACGIH	American Conference of Governmental Industrial Hygienists
AIC	Akaike information criterion
APF	Assigned protection factor
BEI	Biological exposure indices
BZC	Breathing zone concentration
CNV	Copy number variant
CYP450	Cytochrome P450
DNA	Deoxyribonucleic acid
FDR	False discovery rate
GC-MS	Gas-chromatography-mass spectrometry
GFF	Glass-fiber filter
GST	Glutathione-S-transferase
GWAS	Genome-wide association study
HDA	1,6-hexamethylene diamine
HDI	1,6-hexamethylene diisocyanate
HLA	Human leukocyte antigen
LC-MS	Liquid chromatography-mass spectrometry
LD	Linkage disequilibrium
LMM	Linear mixed-effects models
MAF	Minor allele frequency
MDS	Multidimensional scaling
MHC	Major histocompatibility complex
miRNA	Micro RNA
MPP	1-(2-methoxyphenyl)piperazine
NAT	*N*-acetyltransferase
NIOSH	National Institute for Occupational Safety and Health
OSHA	Occupational Safety and Health Administration
PAPR	Powered air-purifying
PBMC	Peripheral blood mononuclear blood cells
PPE	Personal protective equipment
PTFE	Polytetrafluorethylene
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RS	Reference SNP
SNP	Single nucleotide polymorphisms
TLV	Threshold limit values

Introduction

Occupational Exposure Assessment

The science of exposure assessment seeks to measure and monitor exposure to harmful toxicants in order to determine potential risks to human health. This area of research intersects with toxicology, epidemiology, and risk assessment to consider toxic doses in exposed populations, population-wide factors of exposure, and sources and pathways of exposure to determine safe levels of exposure to potentially hazardous chemicals. Hazardous exposures to toxicants are more likely to occur in workplaces and specific high-risk industries. Thus, occupational health scientists often use biological monitoring, i.e., quantifying measurements of metabolites and chemical compounds in biological media such as blood and urine, in workplaces to build predictive exposure models that can be used to assess the risks for development of adverse health effects.

The American Conference of Governmental Industrial Hygienists (ACGIH) publishes a set of threshold limit values (TLV) and biological exposure indices (BEI) that are developed as guidelines to assist in the evaluation and control of workplace health hazards (ACGIH 2016). BEIs are biomarkers or determinants of internal and/or effective dose measured in biological media that can be correlated with levels of chemical exposure and are based on a review of existing peer-reviewed scientific literature by a committee of experts in occupational and public health and related sciences. Regulation of exposures to certain toxic compounds has been federally mandated by the Occupational Safety and Health (OSH) Act of 1970, which ensures that employees work in an environment free from recognized hazards. The OSH Act allows the Occupational Safety and Health Administration (OSHA) to set permissible exposure limits that are federally enforced in workplaces to be protective for the health of workers, while also taking industry interests and economic feasibility into account. Many occupational exposure values were adapted from TLVs and/or recommended exposure limits established by the National Institute for

Occupational Safety and Health (NIOSH), but these values have rarely incorporated genetic variability among workers when setting limits to protect the workforce.

In the last few years with the advent and popularization of high-throughput genotyping technologies, incorporating genomic data into occupational health research has become more feasible. More researchers and regulators have been considering the possibility of using genomic data to develop more tailored risk assessment models for occupational exposures (Christiani et al. 2008; Christiani et al. 2001; Schulte et al. 2015). Variability in acquired genetic effects and inherited genetic make-up can affect toxicodynamic and toxicokinetic processes (Schulte and Howard 2011). In turn, this can differentially modulate internal levels of toxic metabolites and biomarker levels in workers and alter their susceptibility to exposure-induced adverse health effects. Inter-individual variability can also impact the effectiveness of statistical models to predict exposure levels from measured biomarker levels (Figure 1). The state of the science has just started to delve into the rich landscape of using genomic data to inform research in occupational health research.

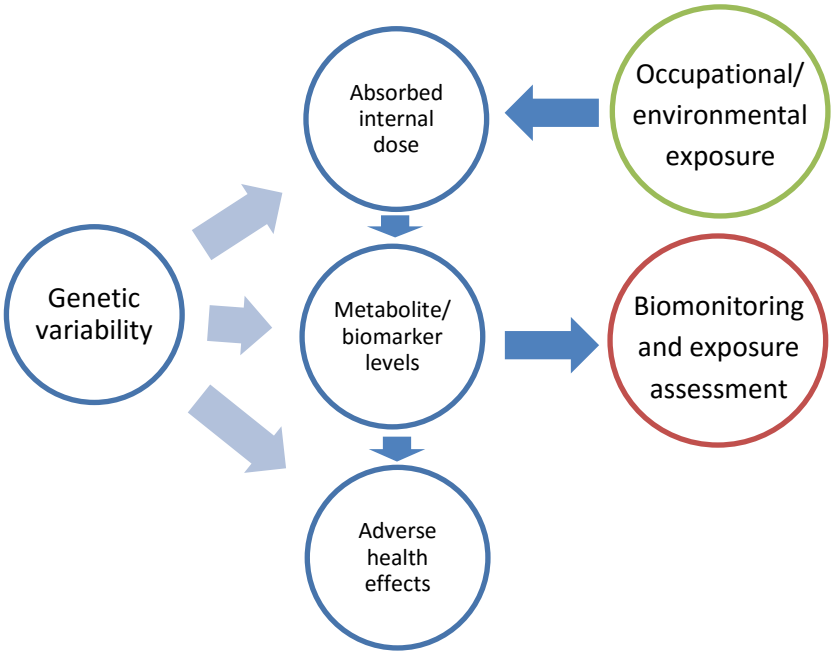
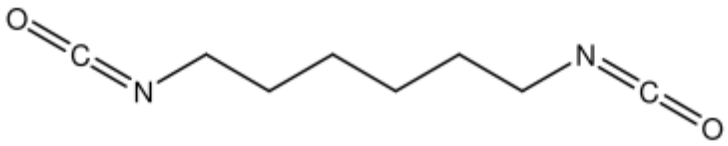


Figure 1. Genetic variability in the context of occupational exposures and evaluation

Diisocyanates

Diisocyanates are a group of low molecular weight organic compounds that are highly reactive due to their two isocyanate functional groups (N=C=O). Common examples include toluene diisocyanate, methylene diphenyl diisocyanate, and 1,6-hexamethylene diisocyanate (HDI; Figure 2).



1,6-Hexamethylene diisocyanate (HDI)
(168.2 amu, 0.05 mmHg)

Figure 2. Molecular structure of 1,6-hexamethylene diisocyanate and its molecular weight and vapor pressure at 25°C

The isocyanate groups are able to undergo exothermic reactions with hydroxyl groups on polyols to form stable and strong polyurethanes (Figure 3). The subsequent products are useful in a variety of industries, and of particular interest to this study, in the application of automotive paint. Diisocyanates are key components of automobile coatings and lacquers, and spray-painters are exposed through inhalation and skin exposure while applying the paints. These compounds are used in these paint formulations to form urethane cross links with alcohols to give the paint a corrosion- and abrasion-resistant finish. Several agencies have set recommended exposure limits for HDI at 0.005 ppm time-weighted average for an 8-hour working day and 0.020 ppm for a short-term 10-minute exposure period based on risks for respiratory irritation and sensitization (ACGIH 2016; NIOSH 2015). This corresponds to a recommended BEI of 15 µg/g creatinine for 1,6-hexamethylene diamine (HDA) in urine measured at the end of shift using acid hydrolysis method (ACGIH 2016).

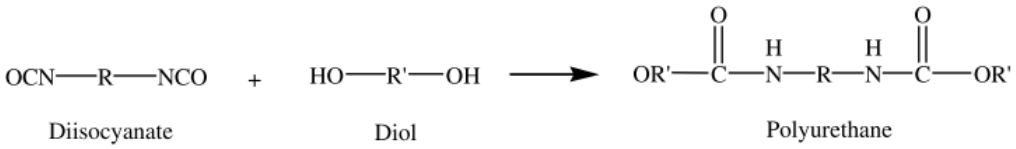


Figure 3. Reaction between diisocyanate and a polyol (diol in this case) to form polyurethane

Health Effects of Diisocyanates

Although diisocyanates are some of the most common chemicals that cause occupational asthma, the mechanism of action for causing allergic reaction and respiratory illness has not been discovered (Piirilä et al. 2000). Immunoglobulin E (IgE) and G (IgG) specific to HDI has been detected in a minority of cases (~20%) which suggests that immunoglobulin mediation may not be the primary mechanism (Budnik et al. 2013; Wisnewski et al. 2012). Studies on diisocyanate-induced asthma are complicated by the long lag-time of sensitization, which can last from months to years, after which point even low exposures to diisocyanates below recommended exposure limits can induce asthmatic attacks (Liu and Wisnewski 2003). As markers of sensitization, there is evidence of systemic immune response in patients who experience diisocyanate-induced asthma because of the presence of diisocyanate-specific antibodies, lymphocyte proliferative responses, and increased cytokine and chemokine production.

Researchers have long harbored suspicions that there are mechanistic and molecular differences between the development of common environmental allergen-induced asthma and diisocyanate-induced asthma (Liu and Wisnewski 2003). Due to the unusual electrophilic reactivity of diisocyanates, the compounds have the ability to directly bind to proteins in exposed cells like those in the airway epithelium, where they can form adducts with proteins such as albumin and glutathione to act as haptens and elicit immune response (Lange et al. 1999). In addition, aided by low molecular weight and bivalent binding sites, diisocyanates form cross-links with various protein species which can then potentially act as carriers for diisocyanate conjugates throughout the body (Wisnewski et al. 1999;

Wisnewski and Redlich 2001). These HDI-conjugated elements form readily and can stimulate lymphocyte proliferation, thus providing a role for HDI-adducts in inducing adverse health outcomes (Wisnewski et al. 1999). Skin exposure may also be involved in the development of diisocyanate sensitivity, though the mechanism by which this exposure route leads to systemic immune response is unclear (Bello et al. 2007; Bello et al. 2006; Herrick et al. 2002).

Diisocyanate-induced asthma is a complex disease with many different protein-protein interactions and molecular pathways at play (Flack et al. 2010a), making it difficult to identify the genetic component of inter-individual variability in the development of disease. Most of the work that has been conducted to determine the gene-environment interactions in asthma and the interplay between epigenetic modifications and exposures in causing disease have focused on environmental allergen-induced asthma, and few research groups have investigated the specific genetic factors associated with occupational asthma (Maestrelli et al. 2009; von Mutius 2009). Lung cytochrome (CYP) P450 enzymes, major histocompatibility complex (MHC) class II genes, human leukocyte antigen (HLA) alleles, CD4 and CD8, *N*-acetyltransferases (NAT1 and NAT2) , interleukins (e.g., IL-13), and variations in glutathione-*S*-transferase (GSTM and GSTT proteins) have been identified with statistically significant associations for increased risk of atopic asthma, allergy, and the development of diisocyanate-induced asthma (Broberg et al. 2008; Liu and Wisnewski 2003; Ober and Hoffjan 2006; Yucesoy et al. 2014).

Though HDI adducts to proteins readily without interacting with metabolic processes, researchers have hypothesized HDI metabolic pathways that could lead to the formation of additional HDI-protein adducts (Figure 4). Also shown in Figure 4 are the various genes that are thought to intersect with these pathways and may impact potential immune response to HDI exposure and elimination routes. Variability in the myriad genes involved in chemical transport, immune response, and any process that affects toxicodynamics and toxicokinetics could impact biomarker levels and exposure assessment.

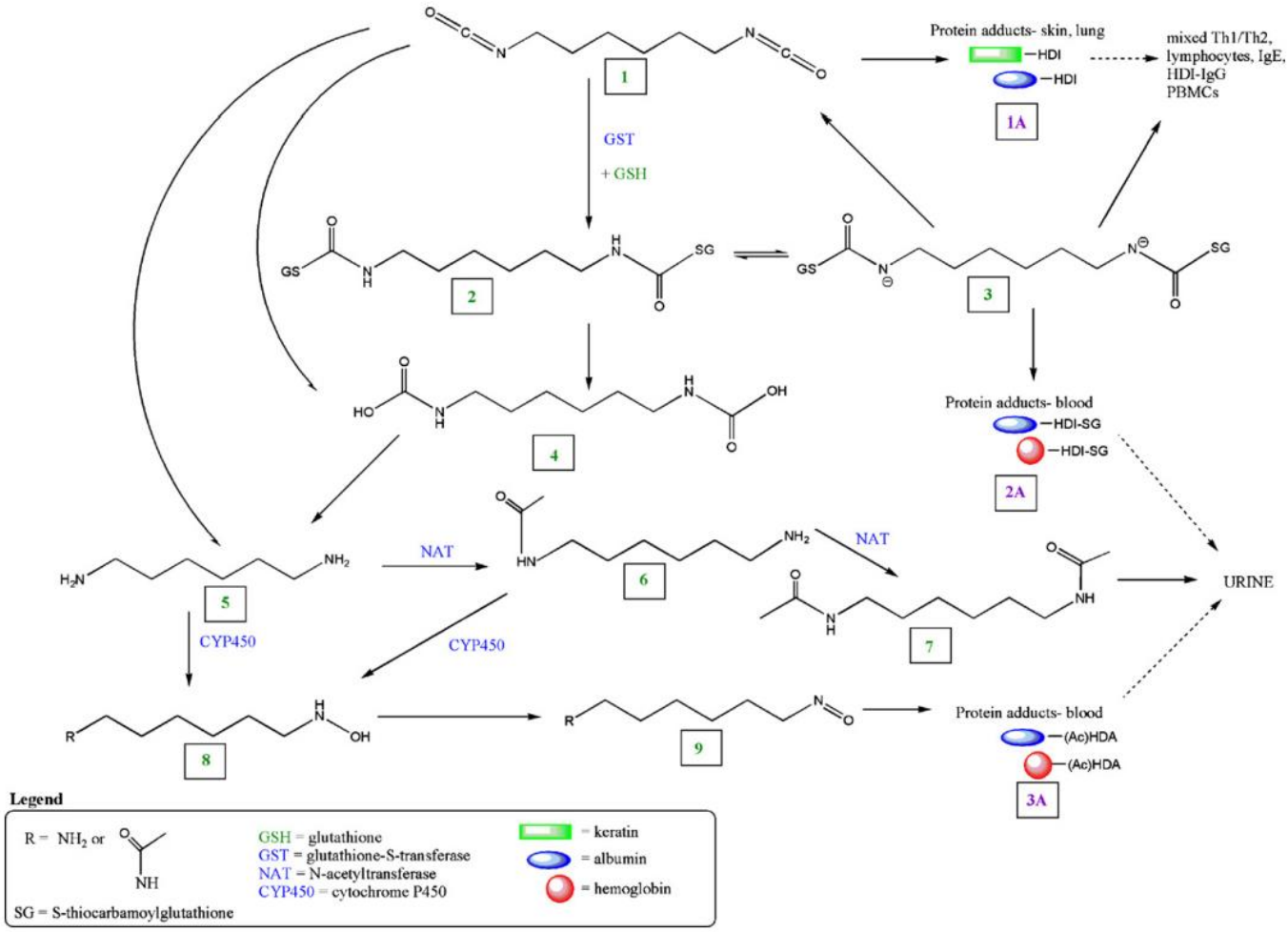


Figure 4. Proposed HDI metabolic pathways (from Flack et al. 2010)

Exposure Monitoring for Diisocyanates

Workers in automobile shops are exposed to HDI when mixing diisocyanate-containing paint and when applying paint with a spray nozzle. The process of spray painting can lead to the aerosolization of paint constituents, leaving behind a vapor or aerosol mist of HDI after the paint has been applied. Consequently, inhalation is a major route of exposure to HDI although there is also evidence that diisocyanates can diffuse through the outermost layer of skin unreacted (Bello et al. 2006; Thomasen and Nylander-French 2012). This exposure occurs even with the use of personal protective equipment (PPE), including ventilated paint booths, respiratory masks, coveralls, and gloves. Our laboratory has

developed standardized and empirically-tested methods to assess breathing-zone concentrations and skin exposure to HDI among auto spray painters exposed on a regular basis (Fent et al. 2009a; Fent et al. 2008; Fent et al. 2006; Fent et al. 2009b; Thomasen et al. 2011a; Thomasen et al. 2011b).

In addition, we have demonstrated the utility of building exposure assessment models with biomarker levels measured in biological media. Acid hydrolysis is the basis of an analytical method for measuring levels of adducts because adding a strong acid to biological samples collected from study participants causes the protein adducts that have formed through direct conjugation or during HDI metabolism to break down and be released in the form of amines (Berode et al. 1991). One such amine, urinary 1,6-hexamethylene diamine (HDA) has been used as a biomarker for systemic exposure to HDI in occupationally-exposed cohorts for several decades (Brorson et al. 1990; Flack et al. 2010a; Gaines et al. 2010a; Gaines et al. 2010b; Gaines et al. 2011). In various exposure assessment models, environmental measurements (i.e., log-transformed breathing-zone concentration of HDI adjusted for the use of respiratory protection and log-transformed skin concentration of HDI) along with several workplace and individual factors (i.e., coverall use, booth type, and log-transformed urinary creatinine concentration) were significant predictors of urine HDA levels (Gaines et al. 2010b; Gaines et al. 2011). Although the parent compound, HDI, is the toxic form of the chemical, HDA is a useful metabolite that can be quantified in exposed individuals to correlate with ambient exposure levels and is thus used in biomonitoring efforts.

Nevertheless, our previous work has also demonstrated that a significant inter-individual variation exists in urine and blood biomarker levels. This variability is not explained by measured HDI monomer and oligomer inhalation and skin exposures alone. In previous work, the final model for predicting log-transformed post-exposure urinary HDA level based on log-transformed breathing zone concentration of HDI, log-transformed creatinine level, and other covariates, indicated that the model explained only 29%

of the variability in the observed biomarker levels (Gaines et al. 2011). In addition, between-worker variation comprised 37.5% of the total observed variance. Genetic heterogeneity among the workers may contribute to a portion of this observed variability in biomarker levels.

Genetics in Occupational Exposure Assessment

Given the complicated molecular underpinnings of diisocyanate-induced sensitization and asthma as well as the wide variability in worker responses to diisocyanate exposure, investigators in recent years have been examining the genetic factors that might explain the development of disease. Among researchers who study diisocyanates, there is a general sense that genetic factors probably explain individual differences in susceptibility to diisocyanate-induced asthma. Only a minority of workers exposed to diisocyanates eventually develop symptoms of diisocyanate asthma (Wisnewski and Redlich 2001). Estimates of the affected workers vary from 5-15% of those exposed due to the difficulty of following up with workers that develop asthma and subsequently leave the industry, known as the "healthy worker effect." Significant associations between the pathogenesis of diisocyanate-induced asthma with several single nucleotide polymorphisms (SNPs) were found in a genome-wide association (GWAS) study that included 74 cases and 824 healthy controls (Yucesoy et al. 2015). The most significant genetic marker among them was rs12913832 located on chromosome 15, which has been mapped to Hepatocyte Nuclear Factor 4 (*HERC2*). In combination with other SNPs discovered through the GWAS, the investigators hypothesized that the identified genes may influence the mechanism of disease through antigen processing and presentation. Previously, an association between SNPs in alpha-T catenin (*CTNNA3*) and diisocyanate-induced asthma was replicated in Korean and Caucasian populations of over 130 affected workers (Bernstein et al. 2013; Kim et al. 2009). These studies suggest that genetic variants play a role in the development of disease, however they generally do not use quantitative biomarker levels or exposure measurements to ascertain actual occupational exposure to diisocyanates.

Another underexplored topic of study is the investigation of potential genetic influences on levels of chemical metabolites measured in the body that are used as biomarkers of exposure. One research group that has investigated the interaction between genetic variability, occupational exposure, and measured biomarker levels found associations between mercury levels in hair, blood, and urine and a panel of 88 SNPs relevant to mercury toxicokinetics in a cohort of 380 dental professionals (Parajuli et al. 2015). Researchers have also had success with finding associations between epigenetic modifications, functional changes in gene expression, and quantitative environmental exposures to arsenic (Rojas et al. 2015). Nylander-French and colleagues have shown that HDI exposure modifies differentially methylated regions across the genome and that these epigenetic modifications partially mediate the HDI exposure and biomarker relationship (Nylander-French et al. 2014). In addition, inherited genetic differences were shown to be statistically associated with adduct levels formed after naphthalene exposure in candidate-gene analysis, suggesting that a quantitative biomarker can be used as an intermediate phenotype when investigating the association between genetic markers and exposure–dose relationship in a small, well-characterized exposed worker population (Jiang et al. 2012).

This report follows up on the progress to date to assess individual genetic variation that may influence the predictive relationship between monitored biomarker levels and measured HDI exposure in the workplace. To date, there has been little work to incorporate individual genetic variability in exposure assessment models. The combination of genetic factors – including allelic differences such as SNPs and other variations in gene expression – and environmental factors has the potential to provide a more complete picture of dose-response relationship to xenobiotic exposure and, thus, more accurate predictive models for assessing exposures to hazardous agents and associated adverse health outcomes in occupational and environmental settings.

Specific Aims

In this present study, we investigated the utility of incorporating genetic variants as covariates in an exposure assessment model for a small worker population of automobile spray painters exposed to 1,6-hexamethyele diisocyanates (HDI). The goals were to:

(1) Identify single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) as markers that are significantly associated with blood and urine biomarkers of HDI exposure;

(2) Incorporate the most significantly associated genetic markers into an exposure assessment model in order to refine its predictive capability.

Methods

Study Population

Characteristics of the participants in this study have been previously described (Flack et al. 2010b; Gaines et al. 2010a). Briefly, automotive repair workers were recruited in central North Carolina and the Puget Sound area of Washington to participate in an assessment of occupational exposures to aliphatic isocyanates. From the total study cohort of 56 spray-painters, 33 workers had complete data on genome-wide markers along with biomarker and exposure measurements. Of these workers, eleven were smokers at the time of data collection, twenty-five identified themselves as non-Hispanic Caucasian, three as Hispanic, one as African-American, one as Asian, one as Native American, and two as mixed race. All subjects were male and ranged in age from 21 to 59 years, with an average age of 35 years. This study was approved by the Institutional Review Board in the Office of Human Research Ethics at the University of North Carolina at Chapel Hill and by the Washington State Institutional Review Board at the Washington State Department of Social and Health Services.

Personal Exposure and Biomarker Measurements

Previous work describes methods of measuring skin and inhalation exposures to HDI (Fent et al. 2008; Gaines et al. 2010a; Gaines et al. 2010b; Gaines et al. 2011), modeling breathing-zone and dermal concentrations of HDI (Fent et al. 2009a; Fent et al. 2009b), and quantification of HDA levels in urine (Gaines et al. 2010a) and blood (Flack et al. 2011; Flack et al. 2010b). To summarize, each of the painters was visited on up to three separate occasions at least one month apart over the course of one year. Personal breathing-zone samples were collected during each spray-painting task in which paint containing monomeric and polymeric HDI was used (e.g., applying surface coating, primer, clear coat). Most of the sampling efforts were focused on clear-coat applications, which comprised 94% of the paint tasks, because that paint formulation contains the highest levels of monomeric and polymeric HDI.

Measurements of HDI in the worker's breathing zone were collected with two-stage filter cassettes (37-mm polystyrene cassette; SKC Inc., Eighty Four, PA) attached to a high-flow air pump operating at 1 L/min. The filter cassette contained an 5-µm pore-size polytetrafluorethylene pre-filter (PTFE; Millipore Corp., Billerica, MA) to collect aerosols and a 1-µm pore-size glass-fiber filter (GFF; SKC Inc.) treated with derivatizing agent, 1-(2-methoxyphenyl)piperazine (MPP) in toluene, to collect vapors. The assigned protection factor (APF) designated by OSHA for the respirator worn by a worker (none, APF=1; air purifying half-face, APF=10; air-purifying full-face, APF=50; supplied air full-face or hood, APF=1000; powered air-purifying (PAPR), full-face or hood, APF=1000) was used to adjust the measured breathing-zone concentrations to account for respiratory protection in inhalation exposure (OSHA 2006). For skin sampling, three consecutive tape strips (4 cm × 2.5 cm, Cover-Roll® adhesive tape Beiersdorf AG, Hamburg, Germany) were collected on the dorsal side of each hand and on the dorsal and volar side of each lower arm immediately after each paint task. The filters and tape strips were then placed into 20-mL glass vials containing derivatizing solution made by dissolving 2 g of MPP in 11 g of 30% v/v solution of *N,N*-dimethylformamide in acetonitrile. Extractions from the filters and tape-strips were analyzed using liquid chromatography-mass spectrometry (LC-MS) with selective ion monitoring as described in Fent et al. (2008).

The collection and analysis of urine HDA and creatinine levels and blood HDA levels have been published previously (Gaines et al. 2010; Flack et al. 2010). Briefly, urine samples were collected at the start of each work day prior to the use of HDI-containing paints and additional spot samples were collected from participating workers each time they urinated. Urine HDA concentration was quantified using gas-chromatography-mass spectrometry (GC-MS) while creatinine concentration was measured using the Creatinine Companion assay kit (Exocell, Inc., Philadelphia, PA, USA). Creatinine was found to be a significant predictor of urine HDA levels in previous studies (Gaines et al. 2010b) so we adjusted the measured urine HDA levels for creatinine in this study as well. End of day blood samples were collected

into heparin and EDTA tubes during each sampling visit when workers consented. Plasma and hemoglobin, extracted from red blood cells, were separated from the whole blood collected in heparin tubes within 24 h of collection, and samples were stored at –40°C until analyzed for HDA concentration by gas chromatography-mass spectrometry (GC-MS) with selective ion monitoring as described in Flack et al. (2010, 2011). The HDA concentrations measured in plasma and hemoglobin were added together as one sum, referred to as the total blood concentration. Questionnaires and work diaries completed at during visit provided information about workers' physical characteristics (e.g., age, height) and workplace factors (e.g., PPE use, paint booth type).

Genotyping

Peripheral blood mononuclear blood cells (PBMCs) were isolated by Ficoll™ separation via centrifugation of whole blood samples collected in EDTA tubes treated with anticoagulant. DNA was purified from PBMC pellets using QiaAmp Blood mini kit (Qiagen, Germantown, MD) and stored in elution buffer at –20°C until analysis of genetic markers. DNA was quantified using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and diluted with 10 mM Tris, pH 7.4 to 50 ng/µL. Genomic DNA isolated from the 33 participants in this study were successfully processed and analyzed on the Affymetrix Genome-wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA) according to the manufacturer's protocol. Qiagen Repli-g genomic amplification kit was used as needed on ten samples to increase yield. The Qiagen kit is designed to provide unbiased and accurate amplification of whole genomes. Some studies have indicated that an appreciable amount of bias is introduced (Pinard et al. 2006), but DNA amplification is recommended by the array manufacturer and is widely conducted. DNA samples were digested with restriction enzymes and purified, and fragments were ligated to adaptors using the Affymetrix SNP 6.0 Core Reagent Kit. A generic primer for the adaptor sequence was used to amplify adaptor-ligated DNA fragments under optimal PCR conditions. These PCR fragments

were pooled, purified, and hybridized to the array, which features ~1.8 million probes, including more than 906,600 SNPs and around 946,000 invariant probes for CNV.

Genetic Array Data Processing and Quality Control

Affymetrix Genotyping Console v4.2 was used to generate genotyping calls and to convert the raw CEL files into PED and MAP files that could be read with PLINK v1.9 (Purcell et al. 2007), an open source genome association analysis toolset. The data were cleaned in PLINK according to standard criteria for quality control (QC). Markers that (1) did not fail the Hardy-Weinberg departure with p-value <0.001, (2) were successfully genotyped at a rate >90%, (3) had a minor allele frequency (MAF) >0.1, and (4) individuals with <10% missing data were included in the subsequent analyses. Only SNPs found on autosomes were included in the analyses because several on the X chromosome did not reliably pass QC. Several iterations of QC were considered, including different levels of MAF (i.e. 0.1, 0.05, and 0.01), and running the process in either PLINK or Affymetrix software. Due of the potential for outsized effects of outliers in our small sample size, setting the MAF criteria at >0.10 was deemed a prudent approach. A comparison of the QC criteria following processing in PLINK and Affymetrix software indicated that PLINK outputted datasets with SNPs passing all of the criteria accurately, with the exception of SNPs on sex chromosomes. Following this exercise, our QC regimens utilized PLINK, set a MAF of >0.10, and only included autosomes in the final dataset.

Candidate-gene Analysis

Candidate genes were determined from the curated published literature that are assumed or known to be involved in the toxicokinetics of HDI or similar chemicals (e.g., toluene diisocyanate and methylene diphenyl diisocyanate) and genes associated with relevant health effects (e.g., occupational asthma) (Broberg et al. 2008; Hoffjan et al. 2003; Liu and Wisnewski 2003; Maestrelli et al. 2009; Mapp et al. 2002; Ober and Hoffjan 2006). PLINK was used to identify SNPs within 20 kb of these specified genes

with locations defined by the human reference genome sequence 18 (NCBI Build 36.1, March 2006). PLINK was used to align the SNPs with proprietary Affymetrix marker IDs and with genotyping data from the Affymetrix array. The selected genes and the number of associated markers are listed in Table 1.

Table 1. Candidate genes tested (N = 19; number of genetic markers = 188)

Gene	Number of Markers	Gene	Number of Markers	Gene	Number of Markers
ADRB2	29	GSTP1	7	IL13	17
CCL5	7	GSTT1	19	LTA	9
CD14	5	HLA-DPB1	11	NAT1	29
CD4	7	HLA-DQA1	38	NAT2	22
CYP1A1	3	HLA-DQB1	35	SERPINA1	8
GSTM1	11	HLA-DRB1	34	TNF	8
GSTM3	21				

Genome-wide Analysis

In addition, a genome-wide analysis (GWAS) was performed using all genotyped markers in exposed workers to identify all genetic variants statistically associated with the biomarker levels (i.e., without a priori evidence of association). Using the cleaned genotyping data from each worker, we used PLINK to determine genetic associations with creatinine-adjusted HDA levels measured in their urine, and with HDA levels measured in plasma, hemoglobin, and total blood (defined here as the sum of HDA measured in plasma and hemoglobin) in separate analyses. The multiple exposure measurements and biomarker levels across work-days were calculated into a single geometric mean value corresponding to each worker in older to simplify regression modeling for significant SNPs in the genetic-association analysis. In exposure assessment studies, quantitative measurements of exposure are often normalized using a logarithmic scale. For the genetic association study, we tested both the geometric means of measurements and log-transformed single-day measurements corresponding to the date of blood draw for DNA collection. The combined, geometric-mean values produced more significant associations between phenotypes and genotypes in our initial analysis. Therefore, geometric mean values of

biomarker and environmental measurements were used as dependent and independent variables, respectively, for all subsequent associations with polymorphisms.

PLINK conducts multiple linear regression to determine associations between quantitative traits, genotypes, and other covariates, with genotypes coded to assume an additive relationship (i.e., AA=0, AT=1, TT=2). Resulting regression coefficients for the SNPs represent the effect of adding each extra minor allele in explaining the biomarker value. Previous research identified covariates significantly associated with urine and blood HDA levels (Flack et al. 2011; Flack et al. 2010b; Gaines et al. 2011) and were incorporated in our model, including: ethnicity, smoking status (both 'past/current smoker' and 'current smoker' statuses were tested), booth type, and coverall and glove usage. Quantitative measurements of HDI exposure in the skin and in the worker's breathing zone adjusted for respiratory protection factor (APF) and paint time were included as covariates in the linear regression in PLINK. Multidimensional scaling (MDS) was performed for cluster analysis of the genotyping data to determine if population substructure may impact the data. However, the first two components of the MDS matrix did not correlate well with self-reported ethnicity so a binary marker of ethnicity (non-Hispanic Caucasian or other ethnicity) was used instead. Using a combination of covariates, 15 models were developed to run in PLINK and compare for associations with biomarker levels (Table 2).

Out of the various combinations, Model 4 captured the greatest number of significant SNPs with overlap of the other models and included (1) current smoking status, (2) ethnicity, as well as (3) the geometric mean values of APF- and paint time-adjusted breathing-zone and (4) skin HDI levels as four separate covariates. Association between a genome-wide selection of SNPs on the Affymetrix array and urine and blood (hemoglobin, plasma, and total blood [hemoglobin + plasma]) biomarker levels was tested using PLINK while controlling for these four covariates.

Table 2: Combinations of binary and quantitative co-variates used in PLINK to determine significant genetic associations with urinary HDA levels.

Model #	**Model Covariates**					
	Smoking status	*HDI breathing-zone exposure*	*HDI skin exposure*	*Ethnicity*	*Coverall use*	*Paint booth type*
1	Current: yes or no (0/1) and ever: yes or no (0/1)	$\mu g/m^3$	ng/mm^3			
2	Current (0/1)	$\mu g/m^3$	ng/mm^3			
3	Ever (0/1)	$\mu g/m^3$	ng/mm^3			
4	Current (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian or other (0/1)		
5	Ever (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)		
6	Current (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)		Type: Down-draft or other (0/1)
7	Ever (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)		Type: (0/1)
8	Current (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)	Usage: Yes or no (0/1)	
9	Ever (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)	Usage (0/1)	
10	Current (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)	Usage (0/1)	Type: (0/1)
11	Ever (0/1)	$\mu g/m^3$	ng/mm^3	Caucasian (0/1)	Usage (0/1)	Type: (0/1)
12	Current (0/1)	$\mu g/m^3$	ng/mm^3		Usage (0/1)	Type: (0/1)
13	Ever (0/1)	$\mu g/m^3$	ng/mm^3		Usage (0/1)	Type: (0/1)
14		$\mu g/m^3$	ng/mm^3			
15	*No covariates included*					

Multiple-testing Correction

The overall type I error rate was controlled by applying false discovery rate (FDR) procedures (Benjamini and Hochberg 1995). This approach seeks to control the expected proportion of false positives, i.e., rejected null hypotheses that should not have been rejected, and in our study, statistical significance was called at FDR $q<0.20$. Proximal SNPs tend to be in linkage disequilibrium (LD) and SNPs in LD may

not be independent from each other, violating independence assumptions for using Bonferonni correction (Gao et al. 2008), leading such a correction to be overly conservative.

Annotation

Affymetrix uses proprietary coding for all of the genetic variants on the array and these were translated to reference SNP cluster IDs (rs numbers) used by dbSNP and many gene ontology programs for further analysis. The annotation file (release 35, 4/30/2015) was downloaded from the Affymetrix website. SAS v9.4 was used to annotate the most significant SNPs in order to perform further bioinformatics analyses and annotations were verified alongside human genome reference sequences 37 and 38.

Gene-ontology Analysis

Significant SNPs from GWAS analysis were compiled and examined for potential network interactions using GeneMANIA (http://genemania.org), a gene ontology enrichment program. GeneMANIA's algorithms and functional interface draws from large datasets of validated protein-protein and protein-DNA interactions and canonical biological pathways to establish predicted networks of interactions and their biological processes or molecular functions (Warde-Farley et al. 2010). These computational tools assess the relative biological plausibility of the genes associated with statistically significant SNPs acting in concert as a cohesive model.

Exposure Models

Linear models were developed in SAS to determine the contribution of genetic markers and combinations thereof to measured biomarker levels using SNPs found to be significantly associated with workers' HDA levels according to the GWAS analysis in urine, plasma, hemoglobin, and total blood. Unlike the genetic association models, these exposure models incorporated repeated or cumulative measurements of occupational exposures to HDI as covariates instead of geometric mean values.

Linear Mixed-Effects Models for Urine HDA Levels

Previous work has shown that urine HDA levels are well-correlated with exposure measurements taken on the same day (Gaines et al. 2010a; Gaines et al. 2010b; Gaines et al. 2011). Therefore, linear mixed-effects models (LMM) were built using PROC MIXED to incorporate repeated measurements corresponding to each worker from multiple sampling visits. Each of the most significant genetic markers was evaluated for association with natural log-transformed urine biomarker levels by fitting the LMM for the i^{th} worker at the j^{th} visit (Y_{ij}), developed to account for random effects (α_i) over multiple collection days. X_{ijp} represents the p^{th} exposure level (inhalation or skin), C_{ijq} represents the q^{th} covariate value (i.e., personal and workplace factors), S_{ig} represents the g^{th} identified genetic marker for the i^{th} worker, and β_p, γ_q, and λ_g represent the corresponding regression coefficients while α_i and ε_{ij}, an independent error term, both have mean 0 and fixed variance. The following LMM was used:

$$\ln(Y_{ij}) = \beta_0 + \sum_{p=1}^{p} \beta_p \ln(X_{ijp}) + \sum_{q=1}^{Q} \gamma_q C_{ijq} + \sum_{g=1}^{G} \lambda_g S_{ig} + \alpha_i + \varepsilon_{ij}$$

The model uses a maximum likelihood-based approach to obtain the most appropriate fit for the observed data and to calculate parameter estimates and associated Wald tests for the fixed effects based on added-last model comparisons (Type III estimates). The random effects in the model represent inter-individual variation for each of the workers between sampling visits. Compound symmetry was chosen as the variance-covariance structure because the workers can all be expected to have similar correlations between separate measurements and variances were modeled as homogenous using this structure. All biomarker and exposure measurements were natural log-transformed to normalize the data. Q-Q plots of the plotted data appeared more normal after this process, indicating that the transformation is appropriate for running regressions using this data, though none of the biomarker values were normally distributed per Shapiro-Wilk tests with $p<0.01$. Comparisons of the Akaike information criterion (AIC) across models was used to determine the most appropriate variables to choose for the model from the significantly-associated SNPs.

Multiple Regression Models for Blood Biomarkers

Blood biomarker HDA levels were better correlated with cumulative HDI exposure than same-day HDI exposure measurements, particularly with adducts of longer half-life such as hemoglobin (Flack et al. 2011). As a result, linear multiple regression models (PROC GLM) were considered for the blood biomarkers with cumulative exposure per worker calculated as a single dependent variable (Y_i) per i^{th} worker. These cumulative exposure estimates were obtained by summing the daily HDI breathing-zone concentration adjusted by paint time and APF or skin concentration across all repeated sampling visits for each worker. Model construction for the linear regression mirrors that of the mixed-effects models, except without random worker effects integrated from the multiple collection visits. All variables were natural log-transformed to meet assumptions of normality and the overall F-value of the model and R^2 values were used to assess the fit of each model.

Regression Diagnostics

Cook's distance (*D*) was calculated in SAS (PROC GLM) in simple ordinary least-squares regression analysis with significant SNPs functioning as independent variables that explain observed biomarker levels. *D* is calculated by the formula shown below, where h_i is the projection matrix with diagonal elements corresponding to the leverage of the i^{th} observation and s^2 is the mean square error of the regression model.

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Multicollinearity was determined by calculation of tolerance ($1\text{-}R^2$) and variance inflation factor (1/tolerance) for each covariate. SAS (PROC SGPLOT) was also used to visualize the distribution of urinary, plasma, hemoglobin, and total blood HDA (geometric mean) by allele (e.g., TT, TA, and AA) via box-and-whisker plots.

Results

Study Population

Table 3 summarizes the characteristics of the workers in this study. We used biological markers of systemic exposure – creatinine adjusted HDA level in urine and HDA levels in plasma and hemoglobin – as intermediate quantitative traits to investigate the association between individual HDI inhalation and skin exposure and exposure-dose response relationship in automobile spray-painters.

Table 3: Summary of the study population characteristics of workers with analyzed genotyping data and complete exposure and biomarker measurement data (n=33)

Variable	Mean	Range
Age (years)	34.8 ± 9.3	21.0 – 59.0
Years in spray-painting job	13.1 ± 10.4	0.5 – 35.0
Number of tasks per worker per sampling visit	2.4 ± 1.3	0 – 5
Time painting during the day (minutes)	21.4 ± 21.0	2 – 98
% of tasks that used clearcoat	97%	
HDI breathing-zone concentration adjusted for respirator APF and paint time (μg/m^3)	14.2 ± 32.8	≤LOD – 257.2
HDI skin concentration (μg/mm^3)	844.1 ± 4160.9	≤LOD – 38087.9
Urine HDA concentration normalized with creatinine (μg/g)	0.79 ± 1.22	≤LOD – 6.45
Plasma HDA concentration (μg/L)	0.09 ± 0.12	≤LOD – 0.71
Hemoglobin HDA concentration (ng/g Hb)	3.77 ± 4.57	≤LOD – 37.19

APF = assigned protection factor; LOD = limit of detection

Genetic Associations with Biomarker Levels

The use of a quantitative dependent variable in statistical analysis increases power when conducting GWAS (Chen et al. 2008). The overall genotyping rate for the worker population was 98.5% and there were 533,673 SNPs that passed quality control. No individuals were discarded due to low genotyping rate. None of the 188 genetic markers assessed in the candidate gene analysis were significantly associated with any of the biomarkers. For the genome-wide analysis, linear regressions run in PLINK produced statistically significant associations for 25 SNPs across the four tested biomarkers ($q<0.20$ after FDR correction). Fourteen SNPs were associated with the natural log-transformed creatinine adjusted

HDA urine levels, 7 SNPs were associated with log-transformed total HDA levels in blood (hemoglobin + plasma), and 4 SNPs were associated with log-transformed plasma HDA levels. Each of these association models incorporated four covariates that included the geometric means of APF- and paint time-adjusted HDI breathing-zone concentration and HDI skin concentration, as well as smoking and ethnicity. The linear models that included the top significant SNPs were reassessed in SAS (PROC GLM) to examine regression diagnostics and to ensure that least-squares assumptions were met. The residuals from the top SNPs (Figure 5) show that they corresponded with expected distributions. Key assumptions of homogeneity of variance and Gaussian distribution of residuals appear satisfied, aside from some deviations at the tails.

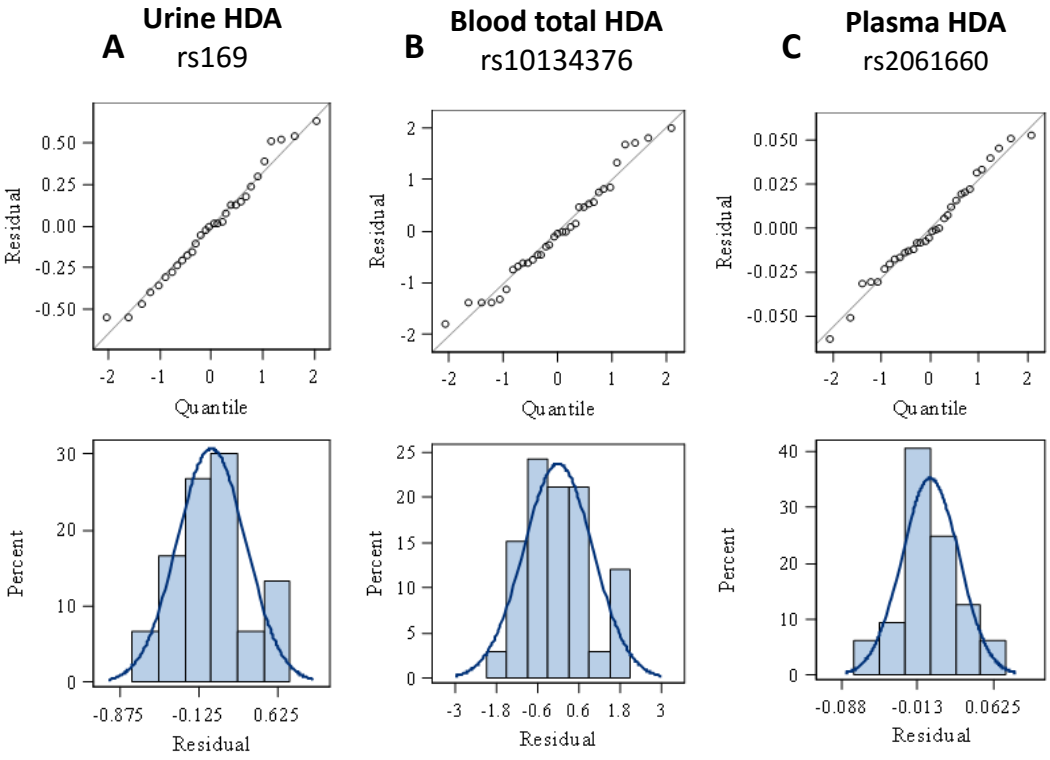


Figure 5: Diagnostic plots from linear models between top SNP associated with each biomarker: (A) geometric mean of urine HDA level adjusted by creatinine, (B) geometric mean of total blood HDA level, and (C) geometric mean of plasma HDA level.

Due to large variance in biomarker levels among the workers in this study, however, the resulting residuals after fitting the linear models with the other SNPs that reached genome-wide significance, particularly with blood total and plasma biomarkers, were not normal even after log-transformation which violates an assumption of conducting least squares regression. Tables 4, 5, and 6 list the SNPs that are statistically associated with the geometric means of urine, blood total, and plasma HDA levels, respectively.

Table 4: Top SNPs significantly associated with geometric mean of total creatinine-adjusted HDA concentration measured in urine.

SNP	P-value	Bonf	FDR	Chr	Position	Alleles	MAF	Associated Gene
rs169	7.9E-08	0.04213	0.03273	7	25046341	C/T	0.2917	OSBPL3 (intergenic)
rs9565949	1.23E-07	0.06546	0.03273	13	85204881	C/T	0.1695	
rs17472697	3.73E-07	0.1992	0.04981	8	29714974	A/G	0.1083	
rs12670377	5.62E-07	0.2998	0.05997	7	96786625	G/T	0.3000	SDHAF3 (intron)
rs9921983	7.79E-07	0.4157	0.06929	16	8376203	A/G	0.1897	
rs7309532	1.02E-06	0.5429	0.07756	12	34483140	A/T	0.2083	
rs489332	1.41E-06	0.7535	0.09419	9	78028346	C/T	0.1780	
rs17692899	2.02E-06	1	0.12	2	29285254	C/T	0.2417	C2orf71 (UTR-3)
rs1343646	2.5E-06	1	0.1334	7	96724452	A/C	0.2667	
rs359250	3.14E-06	1	0.1389	2	60480973	G/T	0.3750	
rs359255	3.14E-06	1	0.1389	2	60484507	A/G	0.3750	
rs6488219	3.43E-06	1	0.1389	12	34460562	A/G	0.2000	
rs400634	3.64E-06	1	0.1389	5	38054753	G/T	0.0678	
rs17685021	4.63E-06	1	0.1646	18	55122775	C/G	0.1667	ONECUT2 (intron)

Bonf: p-value after Bonferroni correction; FDR: False discovery rate q-value; Chr: Chromosome; MAF: minor allele frequency

Table 5: Top SNPs significantly associated with geometric mean of total HDA measured in blood

SNP	P-value	Bonf	FDR	Chr	Position	Alleles	MAF	Associated Gene
rs10134376	1.06E-07	0.05645	0.01882	14	70845874	A/G	0.1333	SYNJ2BP (intron)
rs6573958	1.06E-07	0.05645	0.01882	14	70851036	A/G	0.1333	SYNJ2BP (intron)
rs6573948	1.06E-07	0.05645	0.01882	14	70782154	C/T	0.1333	COX16 (intergenic)
rs7155763	4.06E-07	0.2164	0.05411	14	70846800	C/T	0.1333	SYNJ2BP (intron)
rs6939730	1.34E-06	0.7133	0.1427	6	53071048	A/C	0.1583	LOC105375094
rs8014827	1.96E-06	1	0.1584	14	95541409	C/T	0.1417	DICER1 (intergenic)
rs6575497	2.08E-06	1	0.1584	14	95541198	C/T	0.1441	DICER1 (intergenic)

Bonf: p-value after Bonferroni correction; FDR: False discovery rate q-value; Chr: Chromosome; MAF: minor allele frequency

Table 6: Top SNPs significantly associated with geometric mean of HDA measured in plasma

SNP	P-value	Bonf	FDR	Chr	Position	Alleles	MAF	Associated Gene
rs2061660	3.9E-08	0.02063	0.01031	11	23406939	A/C	0.1333	
rs2061659	3.9E-08	0.02063	0.01031	11	23407014	C/G	0.1333	
rs1454322	2.04E-07	0.1086	0.0362	2	1.59E+08	A/G	0.1167	ACVR1 (intergenic)
rs4870000	6.54E-07	0.3489	0.06979	6	1.52E+08	A/C	0.1083	

Bonf: p-value after Bonferroni correction; FDR: False discovery rate q-value; Chr: Chromosome; MAF: minor allele frequency

Examining the distribution of allele frequencies for these SNPs is illustrative of the means and the ranges of biomarker values across genotypes. Figures 6, 7, and 8 and Tables 7, 8, and 9 show box-and-whisker plots and mean values for each of the most significantly associated SNP with geometric means of urine HDA, blood total HDA, and plasma HDA, respectively. The box displays the 25^{th}, 50^{th}, and 75^{th} quartiles, diamonds mark the mean values, and the whiskers represent the minimum and maximum observations.

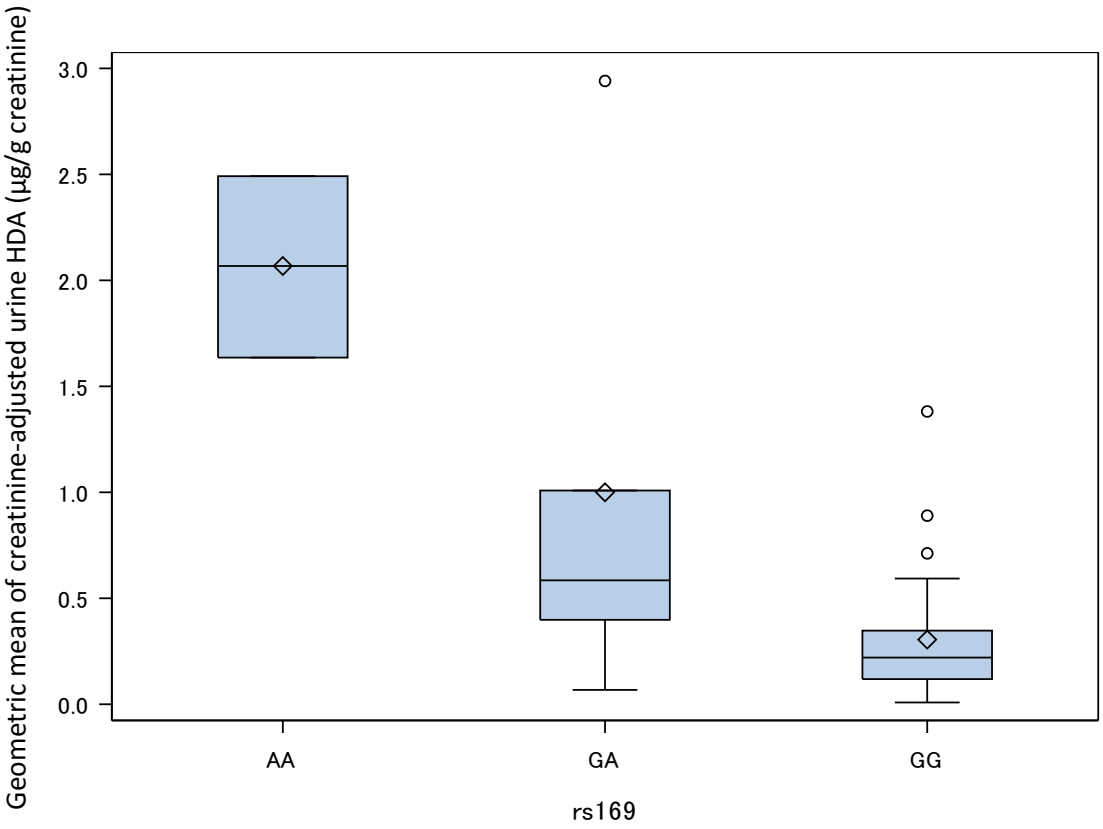


Figure 6. Distribution of allele frequencies for rs169 with geometric mean values of creatinine-adjusted urine HDA concentrations; A is the minor allele.

Table 7. The distribution of creatinine-adjusted urine HDA levels (µg/g creatinine) for each rs169 genotype

rs169	n	Geometric Mean	Geometric Standard Deviation	Minimum	Maximum
AA	2	2.07	0.61	1.64	2.49
GA	5	1.00	1.14	0.072	2.94
GG	23	0.30	0.32	0.011	1.39

n = number of workers

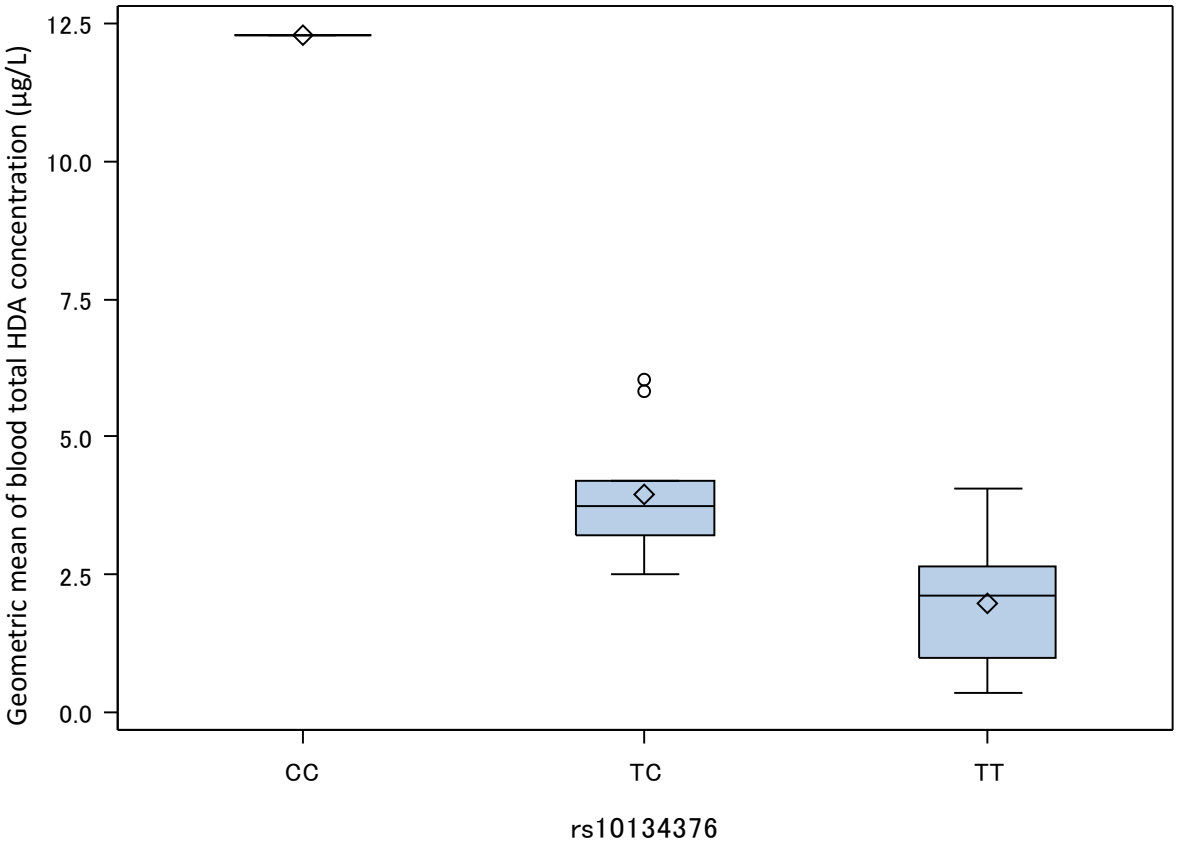


Figure 7. Distribution of allele frequencies for rs10134376 with geometric mean values of blood total HDA concentrations; C is the minor allele

Table 8. The distribution of blood total (plasma + hemoglobin) HDA levels (µg/L) for each rs10134376 genotype

rs10134376	n	Geometric Mean	Geometric Standard Deviation	Minimum	Maximum
CC	1	12.27			
TC	9	3.97	1.21	2.50	6.03
TT	23	1.99	1.05	0.36	4.06

n = number of workers

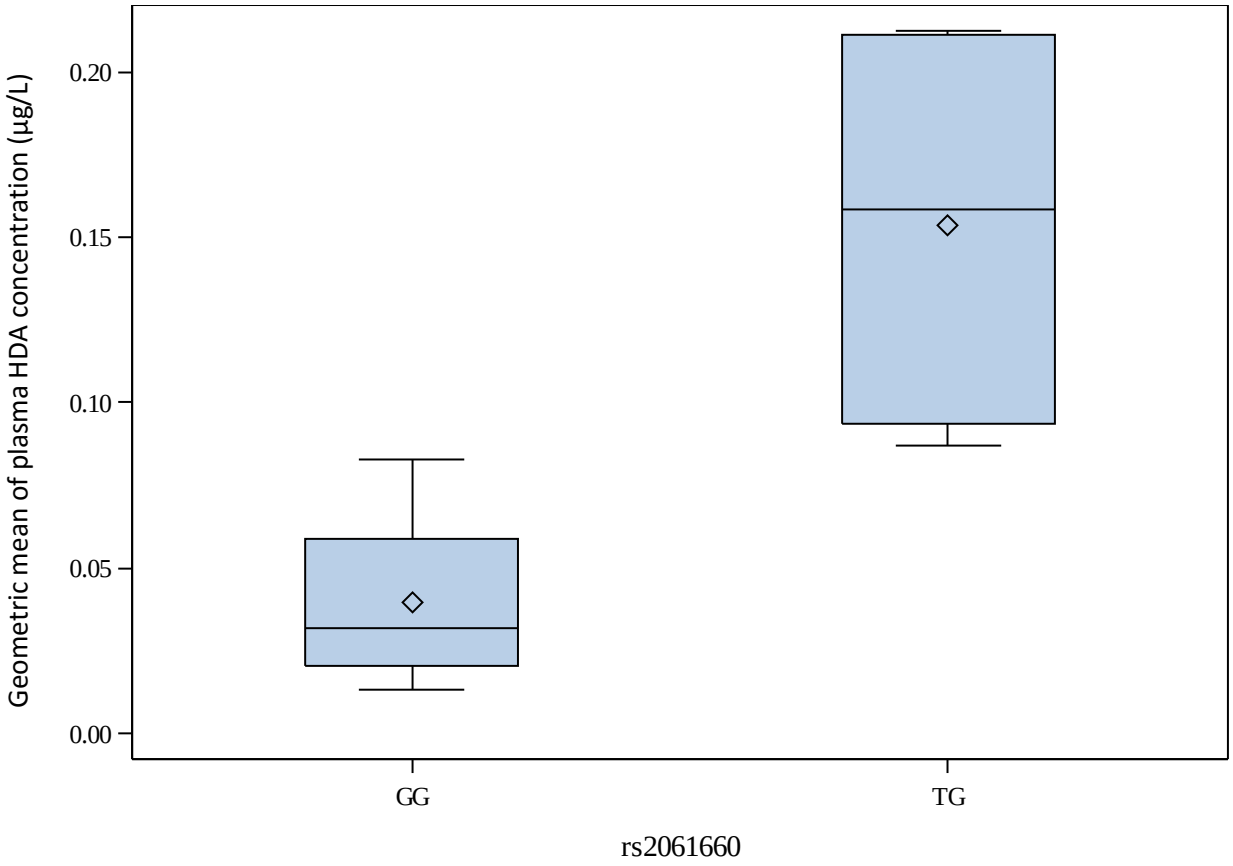


Figure 8. Distribution of allele frequencies of rs2061660 with geometric mean of plasma HDA; T is the minor allele

Table 9. The distribution of plasma HDA levels (µg/L) for each rs2061660 genotype

rs2061660	n	Geometric Mean	Geometric Standard Deviation	Minimum	Maximum
GG	26	0.039	0.022	0.013	0.083
TG	6	0.15	0.055	0.087	0.21

n = number of workers

The use of a quantitative dependent variable relies on several statistical assumptions that drive the appropriateness of using general linear models and least-squares estimation of parameters. The chief difficulty with environmental data is the wide ranges of measurements and the number of non-detected values, resulting in non-homogeneous error variances even after transformations (e.g., natural log-

transformation). As a result, analyses may be skewed by influential measurements and/or subjects with extreme values of variables that are far from the mean in the data. Cook's distance is calculated using values of the independent variables, i.e., HDI exposure measurements in the breathing zone and on the skin, and measures the standardized shift in predicted values and parameter estimates due to the deletion of a particular data point. Cook's distance values (*D*) that are greater than four divided by the number of observations (i.e., D>4/n) is evidence that the data point has high leverage and is influential in determining the slope and intercept of the best-fit model. Omitting these data points in the analysis can produce entirely different results and may indicate that the model fails to capture important characteristics of the model. Cook's distance values from the genetic association between SNPs and blood biomarker levels are shown in Table 10, using the most significant SNP associated with each biomarker in GWAS analysis. In addition, some of the SNPs that reached significant associations with the tested biomarkers had only one individual with an extreme value in the homozygous recessive genotype, which impacts the spread of residuals and the linear fit of the regression. Estimating parameters for groups with only one element is complicated by a lack of variance. Figure 7 shows an example where the recessive genotype has one subject driving the effect and Figure 5 shows the deviation of residuals from predicted values at the tails for all biomarkers.

Table 10: Three highest Cook's distance values for each of the blood biomarkers and corresponding biomarker and exposure measurements

Biomarker	Cook's Distance	GM of biomarker	GM breathing-zone HDI concentration*	GM skin HDI concentration
Blood Total	3.89	2.73	184.8	212.2
Blood Total	0.19	6.03	4.07	0.001
Blood Total	0.10	4.06	38.8	423.2
Plasma	10.08	0.043	184.8	212.2
Plasma	0.50	0.094	6.82	2583.9
Plasma	0.40	0.064	35.6	3066.6

GM=geometric mean; *Breathing zone HDI concentration is APF and paint time-adjusted

Exposure Assessment Models

Urine Biomarkers

Data from multiple visits was used to generate a LMM wherein the most significant SNP associated with urine HDA level, rs169, and log-transformed skin HDI exposure were significant predictors of measured biomarker levels. Each of the LMMs of the form: $\ln(Y_{ij}) = \beta_0 + \sum_{p=1}^{2} \beta_p \ln(X_{ijp}) + S_i + \alpha_i + \varepsilon_{ij}$, that individually incorporated the top five SNPs significantly associated with the urine HDA level showed that each SNP (*S*) is a significant predictor in its respective model. Log-transformed creatinine-adjusted urine HDA level was the dependent variable (*Y*) in each of these models and log-transformed skin HDI and APF- and paint-time adjusted breathing-zone HDI concentration were included in the model as well covariates (*X*'s). The model that included rs169 as a covariate had the lowest AIC=318.9 compared with models that included the next four SNPs: rs9565949, rs17472697, rs12670377, and rs9921983. These models that incorporated each subsequent SNP individually had AIC values of 325.9, 340.0, 341.1, and 353.6, respectively.

Table 11. Solutions for fixed effects from linear mixed model (Wald tests) incorporating the most significant SNP rs169 and exposure measurements; dependent variable is natural log-transformed creatinine adjusted urine HDA

Effect	rs169	Estimate	Standard Error	DF	t	p-value
Intercept		-1.73	0.266	27	-6.51	<0.0001
Log-BZC HDI		0.059	0.093	48	0.63	0.530
Log-skin HDI		0.075	0.036	48	2.09	0.042
rs169	AA	2.11	0.845	27	2.50	0.019
rs169	GA	0.815	0.611	27	1.33	0.193
rs169	GG					*reference*

Log-BZC HDI = Log-transformed HDI breathing-zone concentration adjusted for respirator assigned protection factor and paint time (μg/m^3); Log-skin HDI = Log-transformed HDI skin concentration (μg/mm^3)

Multicollinearity among the covariates was not an issue as each of the variance inflation factor values calculated for the exposure measurements of HDI were near 1 and correlations between the SNPs were $\rho<0.75$. Solutions for the fixed effects from the added-last Wald tests are summarized in Table 11.

Blood Biomarkers

We also developed multiple linear regression models that tested the association between the cumulative blood biomarker levels measured in each worker with the measured cumulative HDI inhalation and skin exposures and incorporated the impact of selected SNPs. Eigenanalysis and review of the correlation matrix showed almost complete agreement ($\rho=0.9$-1) between the most significant SNPs associated with plasma and complete correlation ($\rho=1$) between the most significant SNPs associated with total blood HDA levels. Therefore, linear models could only be developed with one SNP at a time, making these models similar to those in the initial association step that determined SNPs that were significantly associated with the geometric-mean of biomarker levels. While the Wald *t*-values, summarized in Table 12, indicate that each SNP is a significant predictor ($p<0.05$) of cumulative blood biomarker levels after taking into account cumulative HDI exposure measurements, the diagnostic plots appear less robust for least-squares assumptions compared with initial association models (Figure 9).

Table 12. Solutions for added last Wald tests from linear regression model incorporating most significant SNP and exposure measurements; dependent variables are (A) cumulative natural log-transformed total blood HDA and (B) cumulative natural log-transformed plasma HDA

A

R^2=0.37

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rs10134376	2	37.6	18.8	5.99	0.0068
Cumulative Log-BZC HDI	1	4.87	4.87	1.55	0.2232
Cumulative Log-Skin HDI	1	0.597	0.597	0.19	0.6662

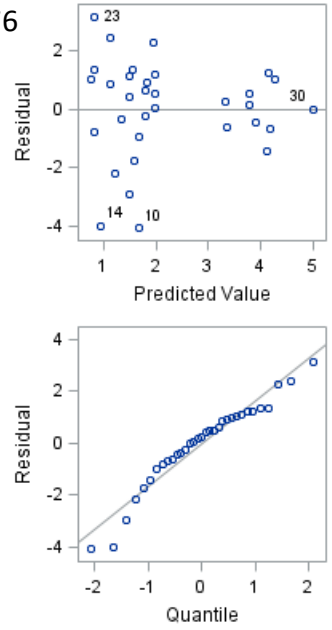
B

R^2=0.53

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rs2061660	1	57.4	57.4	14.78	0.0006
Cumulative Log-BZC HDI	1	2.66	2.66	0.68	0.4149
Cumulative Log-Skin HDI	1	7.17	7.17	1.85	0.1851

Log-BZC HDI = Log-transformed HDI breathing-zone concentration adjusted for assigned protection factor and paint time (µg/m^3); Log-skin HDI = Log-transformed HDI skin concentration (µg/m

A. Blood total HDA
rs10134376



B. Plasma HDA
rs2061660

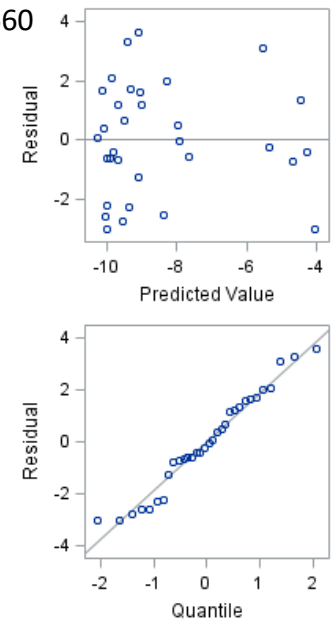


Figure 9: Diagnostic plots from model with most significant SNP associated with each biomarker: (A) cumulative log-transformed total blood HDA, and (B) cumulative log-transformed plasma HDA.

Bioinformatics

Urine Biomarkers

Significant SNPs associated with urine HDA levels are in turn associated with the following genes: *OSBPL3*, *ACN9* (also annotated as *SDHAF3*), *C7orf31*, and *ONECUT2* (Table 4). These query genes were analyzed using GeneMANIA, gene ontology software that finds a small set of genes that are most likely to share function with that gene based on their interactions using data from hundreds of functional genomics datasets. The predicted gene networks of physical and genetic interactions, shared protein domains, co-expression, etc., are summarized in Figure 10. The *OSBPL* gene family is thought to comprise proteins that bind oxysterol and form a group of intracellular lipid receptors (Weber-Boyvat et al. 2013). This family is involved in the regulation of cell adhesion and the organization of the actin cytoskeleton, and may be important for the transport of the compound in relation to epithelial cells.

The 22 genes that have demonstrated interactions or network intersections with three of the query genes have known functions that are listed in Table 13. *C7orf31* has no known molecular functions at this time and was not informative to the bioinformatics analyses. These four pathways - sterol, alcohol, steroid, and cholesterol binding - are enriched because the predicted network interactions have more participation in these select networks than would be expected due to random chance.

Table 13. Predicted molecular functions of genes with known interactions associated with urine HDA.

Function	False discovery rate	Coverage
Sterol binding	1.25e-10	7/34
Alcohol binding	1.33e-9	7/51
Steroid binding	1.36e-9	7/54
Cholesterol binding	3.94e-9	6/30

Blood Biomarkers

In addition, the following genes are associated with SNPs found to be significantly associated with blood total blood HDA (hemoglobin + plasma): *COX16*, *SYNJ2BP*, and *DICER1* (Table 5).

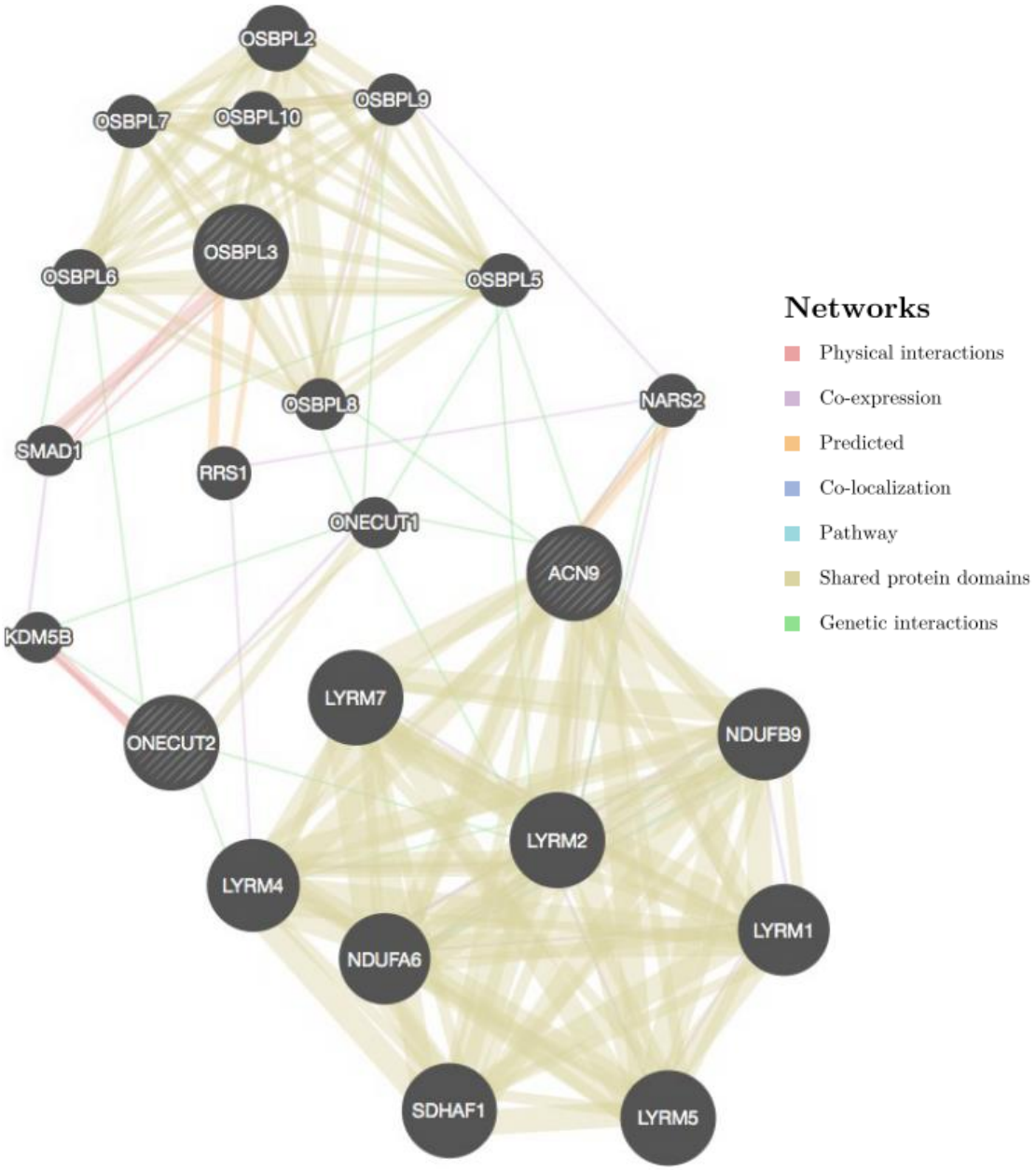


Figure 10. Predicted network interactions based on enrichment of molecular functions derived from three candidate genes associated with log-transformed creatinine adjusted urine HDA levels.

GeneMANIA was again used to assess putative gene pathways that interact with these query genes (Figure 11). The SNPs that are associated with either *SYNJ2BP* or *COX16* are part of a *SYNJ2BP-COX16* fusion transcript and are located in a CNV rich locus. The function of the genes that have demonstrated interactions with the query genes (Figure 11) are implicated in the pathways shown in Table 14.

Table 14. Predicted molecular functions of genes with known interactions associated with HDA blood biomarkers (total blood)

Function	FDR	Coverage
Gene silencing by RNA	1.60e-17	10/42
Gene silencing	1.56-15	10/68
Posttranscriptional gene silencing by RNA	3.14e-11	7/34
Posttranscriptional gene silencing	3.14e-11	7/34
Regulation of gene expression, epigenetic	3.74e-11	9/123
Gene silencing by miRNA	3.25e-9	6/31

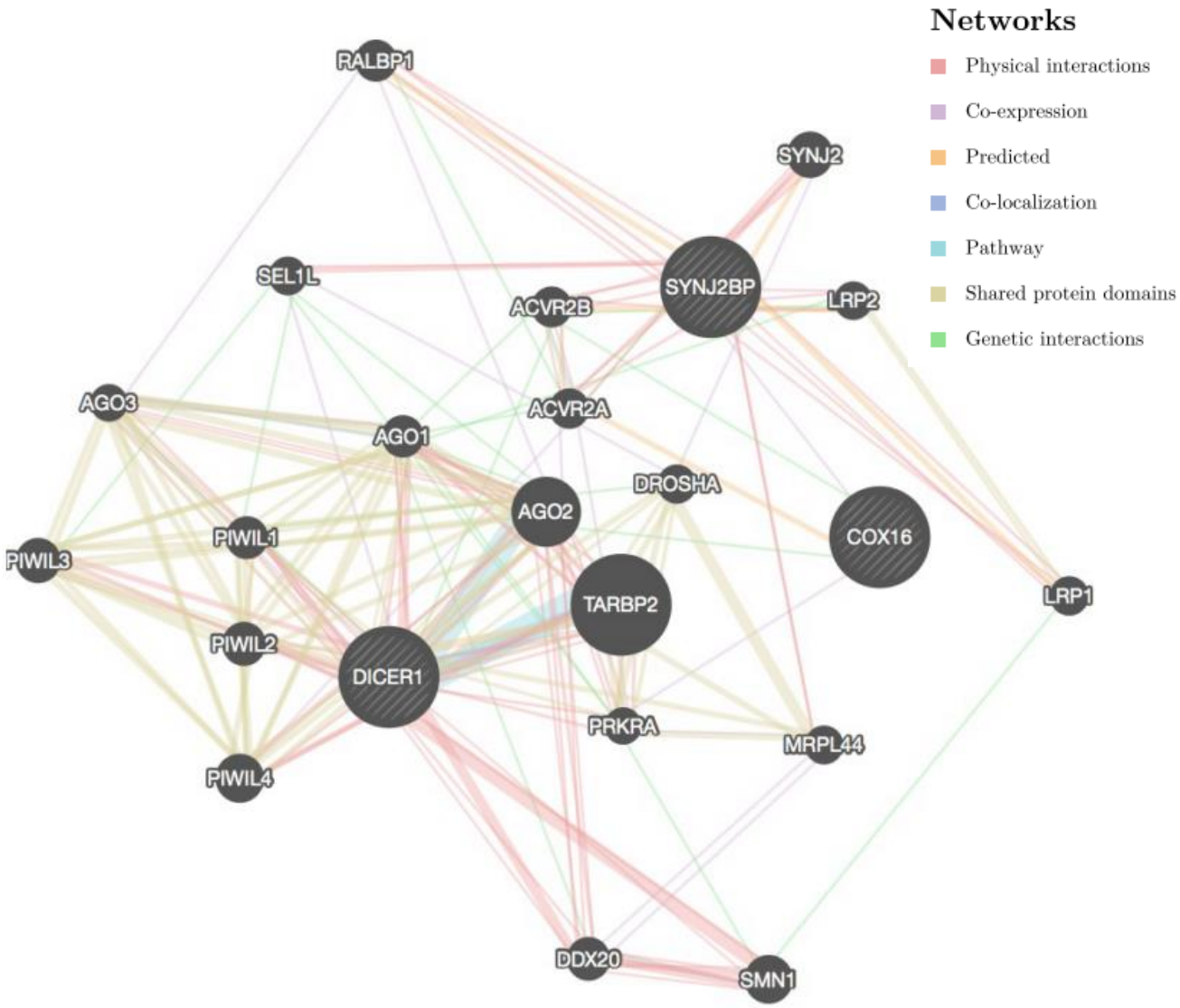


Figure 11. Predicted network interactions based on enrichment of molecular functions derived from three candidate genes associated with log-transformed total blood HDA levels.

Discussion

Obtaining accurate quantitative data on environmental exposures strengthens the applicability of toxicological studies and biomonitoring efforts to estimate actual risks for the development of human health effects. This present study builds upon a small and well-characterized occupational cohort that has been used successfully to assess exposure-dose relationships in HDI exposed workers and to inform exposure assessment models (Fent et al. 2009a; Fent et al. 2008; Fent et al. 2006; Fent et al. 2009b; Flack et al. 2008; Flack et al. 2010a; Flack et al. 2011; Flack et al. 2010b; Gaines et al. 2010a; Gaines et al. 2010b; Gaines et al. 2011; Thomasen et al. 2011a; Thomasen et al. 2011b; Thomasen and Nylander-French 2012). Workers in our study population (n=33) were genotyped and the data were used to discover SNPs that are statistically associated with levels of biomarkers measured in biological media from workers exposed to HDI. We report on the feasibility of determining statistically significant associations between HDA biomarker levels and markers of genetic variability with three types of biomarkers – urine HDA level adjusted for creatinine level, plasma HDA level, and total blood HDA level (plasma + hemoglobin). Furthermore, we found that the associated SNPs are then significant predictors in exposure models built upon repeated measurements (for urine HDA) and cumulative measurements (for blood biomarkers) of biomarker and exposure levels to HDI.

SNP rs169 was significantly associated with the geometric mean of creatinine-adjusted urine biomarker level in an initial linear regression model that included four covariates of exposure (breathing-zone concentration of HDI adjusted by APF and paint time, skin concentration of HDI, ethnicity, and smoking). In a separate LMM conducted with PROC MIXED that incorporated repeated exposure measurements of HDI as covariates, SNP rs169 was again a significant predictor of urine HDA adjusted by creatinine measured over multiple visits per worker. A total of 14 SNPs were found to be significantly associated with urine HDA holding FDR at $q<0.20$ and each of the top five SNPs were found to be significant

predictors in linear mixed effects models when run individually. The model that contained rs169 was the best-fit model as determined by AIC compared with models that incorporated the other SNPs. Ideally, the exposure assessment model would have incorporated multiple SNPs to better account for inter-individual variability, but in LMMs that included multiple SNPs as covariates for association with urine biomarker levels, no SNPs were observed to be significant predictors.

A total of 7 SNPs were significantly associated with the geometric mean of total blood HDA (calculated as a sum of plasma HDA and hemoglobin HDA) and 4 SNPs were significantly associated with the geometric mean of plasma HDA after FDR correction. Of those, rs10134376 was the most significantly associated SNP with total blood HDA and rs2061660 was the most significantly associated SNP with plasma HDA. Each of these SNPs was also a significant predictor in the linear regression models that assessed the association between cumulative blood biomarkers and cumulative breathing-zone and skin concentrations of HDI. The top SNPs associated with total blood HDA were highly correlated and there is evidence that the markers are in linkage disequilibrium (LD) with each other because their chromosomal positions on chromosome 14 are closely spaced together. Similarly, several of the top SNPs associated with plasma HDA appear to be in LD. Therefore, only linear models that incorporated one SNP at a time could be assessed due to the presence of collinearity between the SNPs.

The present method is sensitive to statistical outliers in the data with high values of the dependent and independent variables. The SNPs that are determined to be significantly associated with biomarker values differ widely under various association criteria, indicating that the GWAS analysis may not identify stable markers of individual susceptibility in our small cohort. Genotypes corresponding to the individuals with particularly high or low biomarker values can produce high residuals and impact the appropriateness of fitting linear models with the data and complicate the estimation of parameters. In addition, variation in the environmental measurements of HDI can also have influence on the

association model as measured by Cook's distance (e.g., the worker homozygous for the minor allele of the SNPs most associated with total blood and plasma). The categorization of these individuals as influential points is not necessarily a problem for exposure assessment, in general, because such high exposure and biomarker levels can and do occur in the workplace. However, for the purposes of statistical modeling and determining genetic associations, an outlier in the data impacts the distribution of residuals from the fitted model and may impact the validity of positive associations. Changes in the exact association criteria used for the GWAS analysis, e.g., allele coding and minor allele frequency criteria, can have the impact of altering the number and order of significant SNPs. As such, the utility of the linear exposure assessment models produced in this study should be validated by a separate set of samples to ensure replicability and broad applicability.

The AA genotype of rs169, which is homozygous for the minor allele, is associated in the linear mixed-effects model with a large increase of creatinine-adjusted urine HDA level. The parameter estimate for the recessive genotype dwarfs the biomarker measurements for most workers in the cohort, which may also explain why the environmental measurements, i.e., log-transformed breathing-zone HDI and log-transformed skin HDI concentration, that were found to be predictive of urine and blood biomarker levels in previous studies were not significant in this analysis (Flack et al. 2010a; Flack et al. 2011; Flack et al. 2010b; Gaines et al. 2010b; Gaines et al. 2011). The fixed effects shown in Table 11 are Wald estimates from added-last tests that interrogate the added value of a particular covariate on top of a model with all of the other covariates already in place. Because the estimated effect-size of the SNPs in the linear mixed effects models exceeds the effect-sizes for the breathing-zone concentration of HDI and the skin concentration of HDI, the addition of these estimates in a model that already includes the SNPs may be less likely to have significant predictive effects.

The use of a quantitative dependent variable in the genetic association study may have been a double-edged sword in the case of this dataset. On one hand, the wide range in biomarker levels among the workers provided potentially high effect-sizes that could identify SNPs with significant associations. Dichotomization of disease endpoints as case or control leads to a loss of information which can be critical in assessing association effects, particularly with exposure studies. However, in our specific analysis and with our small cohort, the range in biomarker and exposure measurements could also present a source of bias in the data that skews the least squares regression modeling in a direction that is not representative of each SNP's actual biological influence on biomarker levels. In addition, GWAS often suffers from the "winner's curse" in which the effect-sizes of variants may be overestimated in initial, exploratory results because scientists tend to focus only on variants that yield significant evidence for association (Xiao and Boehnke 2011).

Bioinformatics

Genes that are directly associated with the biomarkers that we tested, i.e., *DICER1, COX16, SYNJ2BP, OSBPL3, ACN9*, and genes that are known to interact in the same networks as these genes (Figure 10 and 11) are implicated in various pathways. In particular, several of these genes are involved in binding pathways, which may play a role in the transport of xenobiotics and their metabolites and other elements that may have an impact on the levels of biomarkers detected in blood and urine. Members of the *OSBPL* gene family are thought to be proteins that bind oxysterol and form a group of intracellular lipid receptors (Weber-Boyvat et al. 2013). This family is involved in the regulation of cell adhesion and the organization of the actin cytoskeleton.

Mutations in *DICER1* and *TARBP2* have been linked with impacts on gene silencing and micro RNA (miRNA) processing due to the key role that these proteins have in the RNA-induced silencing complex (RISC) that helps load miRNA in the process of RNA interference (Daniels et al. 2009; Tijsterman and

Plasterk 2004). This could have implications on gene expression and other alterations that may mediate the effects between environmental exposures and molecular responses. In addition, the locus on chromosome 14 associated with several SNPs in our analysis that are statistically significant markers of blood total HDA levels represents naturally occurring read-through transcription between the neighboring *SYNJ2BP* (synaptojanin 2 binding protein) and *COX16* (COX16 cytochrome c oxidase assembly homolog (S. cerevisiae)) genes. The read-through transcript produces a fusion protein that shares sequence identity with each individual gene product. Alternate splicing results in multiple transcript variants that encode different isoforms. In a recent, large meta-analysis of GWAS data from well-established research consortia, polymorphisms in this region have been associated with a decrease in systolic blood pressure. Because these SNPs were associated with total blood levels of HDA, this connection could illustrate the implications of inter-individual differences in blood pressure, and conceivably, the impact of metabolite and adduct transport on biomarker levels (Simino et al. 2014).

Bioinformatics analyses using the agnostic approach based on whole genome protein-protein or protein-DNA interaction rather than exclusively curated literature interaction provide further jumping off points for hypothesis generation and suggest mechanistic pathways for functional analysis (reverse genetics) that may explain the phenotypes and biomarker levels that we measured. In subsequent studies it is pertinent that we understand the potential contribution, if any, of these genetic markers to the variability in biomarker levels and exposure assessment. Ultimately, significantly associated polymorphic genetic variants will require functional validation using molecular biology and reverse genetic techniques in appropriate cell based or tissue reconstructs for confirmation of their role in modification of the biomarker of exposure levels.

Genetics in Occupational Exposure Assessment

In response to the recent attention devoted to genetics, scientists have recommended minimum criteria necessary for chemical-specific analysis of the effect of a polymorphism on tissue dose:

> 1. Well-characterized metabolic pathway, with relevant isozyme identified for all major steps;
> 2. Allelic frequency data available for all major polymorphic enzymes,
> 3. Phenotype data for the chemical of interest for each major variant allele, and
> 4. Existing physiologically based pharmacokinetic (PBPK) model or development of an adequate model to describe polymorphism data.
>
> (Gentry et al. 2002)

Toxicological data on the mechanisms and modes-of-action for many chemicals used in workplaces falls short of these standards. In particular, little is known about the metabolic pathways for diisocyanates in humans and less still is known about the heritability and penetrance of the genetic variability that may affect these pathways or others that may alter biomarker levels, e.g., blood pressure, kidney function, or other physiological effects that maintain homeostasis. Without such prior knowledge, it is difficult to determine the biological mechanisms that explain any positive associations found in genome-wide studies. The influx of genetic data in various fields of health sciences research opens exciting possibilities for more personalized estimates on health metrics, but it also requires a thoughtful approach to ensure this data is useful.

Genomic data could hold great value for occupational health research as well by helping to identify susceptible subpopulations for certain workplace exposures. However, as of now, genetic screening for susceptibility factors related to the adverse health effects of toxicant exposures is not widely conducted. On a large scale, genetic screening for employment and workplace safety is not necessarily recommended as this could result in wide-reaching ethical and social concerns. In addition, the positive

predictive value of genetic screening is weak even for well-established links such as the *HLA-DPB1* marker of susceptibility to chronic beryllium disease among workers exposed to beryllium (Christiani et al. 2008; Christiani et al. 2001). Genetic association studies of diisocyanate-induced occupational asthma have found inconsistent effects of the distribution of HLA class II alleles in Italian and American cohorts (Beghé et al. 2004). Nevertheless, investigators are optimistic that genetic variability can help identify high-risk groups and guide further research on inter-individual variability in responses to toxicant exposures (Dix et al. 2006; Schulte et al. 2015).

In the present study, we find that genetic variability can be incorporated in exposure assessment models that are predictive for measured biomarker levels. However, due to variability in environmental and biomarker measurements and the small size of our study cohort, it is difficult to ascertain whether or not our findings are specific to spray painters occupationally exposed to HDI. The linear models that we have developed are disproportionately impacted by a few people in the study, so the estimates may not be representative of the wider worker population of individuals impacted by exposure to diisocyanates. More robust data needs to be collected, and particularly in larger sample sizes, to verify GWA findings and to provide a mechanistic basis for the increased susceptibility of certain individuals.

References

ACGIH. 2016. 2016 tlvs and beis: 7th edition documentation.

Beghé B, Padoan M, Moss CT, Barton SJ, Holloway JW, Holgate ST, Howell WM, Mapp CE. 2004. Lack of association of hla class i genes and tnf alpha-308 polymorphism in toluene diisocyanate-induced asthma. Allergy. 59(1):61-64.

Bello D, Herrick CA, Smith TJ, Woskie SR, Streicher RP, Cullen MR, Liu Y, Redlich CA. 2007. Skin exposure to isocyanates: Reasons for concern. Environ Health Perspect. 115(3):328-335.

Bello D, Smith TJ, Woskie SR, Streicher RP, Boeniger MF, Redlich CA, Liu Y. 2006. An ftir investigation of isocyanate skin absorption using in vitro guinea pig skin. J Environ Monit. 8(5):523-529.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 57(1):289-300.

Bernstein DI, Kashon M, Lummus ZL, Johnson VJ, Fluharty K, Gautrin D, Malo JL, Cartier A, Boulet LP, Sastre J et al. 2013. Ctnna3 (α-catenin) gene variants are associated with diisocyanate asthma: A replication study in a caucasian worker population. Toxicol Sci. 131(1):242-246.

Berode M, Testa B, Savolainen H. 1991. Bicarbonate-catalyzed hydrolysis of hexamethylene diisocyanate to 1,6-diaminohexane. Toxicol Lett. 56(1-2):173-178.

Broberg K, Tinnerberg H, Axmon A, Warholm M, Rannug A, Littorin M. 2008. Influence of genetic factors on toluene diisocyanate-related symptoms: Evidence from a cross-sectional study. Environ Health. 7:15.

Brorson T, Skarping G, Nielsen J. 1990. Biological monitoring of isocyanates and related amines. Ii. Test chamber exposure of humans to 1,6-hexamethylene diisocyanate (hdi). Int Arch Occup Environ Health. 62(5):385-389.

Budnik LT, Preisser AM, Permentier H, Baur X. 2013. Is specific ige antibody analysis feasible for the diagnosis of methylenediphenyl diisocyanate-induced occupational asthma? Int Arch Occup Environ Health. 86(4):417-430.

Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK et al. 2008. Variations in dna elucidate molecular networks that cause disease. Nature. 452(7186):429-435.

Christiani DC, Mehta AJ, Yu CL. 2008. Genetic susceptibility to occupational exposures. Occup Environ Med. 65(6):430-436; quiz 436, 397.

Christiani DC, Sharp RR, Collman GW, Suk WA. 2001. Applying genomic technologies in environmental health research: Challenges and opportunities. J Occup Environ Med. 43(6):526-533.

Daniels SM, Melendez-Peña CE, Scarborough RJ, Daher A, Christensen HS, El Far M, Purcell DF, Lainé S, Gatignol A. 2009. Characterization of the trbp domain required for dicer interaction and function in rna interference. BMC Mol Biol. 10:38.

Dix DJ, Gallagher K, Benson WH, Groskinsky BL, McClintock JT, Dearfield KL, Farland WH. 2006. A framework for the use of genomics data at the epa. Nat Biotechnol. 24(9):1108-1111.

Fent KW, Gaines LG, Thomasen JM, Flack SL, Ding K, Herring AH, Whittaker SG, Nylander-French LA. 2009a. Quantification and statistical modeling--part i: Breathing-zone concentrations of monomeric and polymeric 1,6-hexamethylene diisocyanate. Ann Occup Hyg. 53(7):677-689.

Fent KW, Jayaraj K, Ball LM, Nylander-French LA. 2008. Quantitative monitoring of dermal and inhalation exposure to 1,6-hexamethylene diisocyanate monomer and oligomers. J Environ Monit. 10(4):500-507.

Fent KW, Jayaraj K, Gold A, Ball LM, Nylander-French LA. 2006. Tape-strip sampling for measuring dermal exposure to 1,6-hexamethylene diisocyanate. Scand J Work Environ Health. 32(3):225-240.

Fent KW, Trelles Gaines LG, Thomasen JM, Flack SL, Ding K, Herring AH, Whittaker SG, Nylander-French LA. 2009b. Quantification and statistical modeling--part ii: Dermal concentrations of monomeric and polymeric 1,6-hexamethylene diisocyanate. Ann Occup Hyg. 53(7):691-702.

Flack S, Goktepe I, Ball LM, Nylander-French LA. 2008. Development and application of quantitative methods for monitoring dermal and inhalation exposure to propiconazole. J Environ Monit. 10(3):336-344.

Flack SL, Ball LM, Nylander-French LA. 2010a. Occupational exposure to hdi: Progress and challenges in biomarker analysis. J Chromatogr B Analyt Technol Biomed Life Sci. 878(27):2635-2642.

Flack SL, Fent KW, Gaines LG, Thomasen JM, Whittaker SG, Ball LM, Nylander-French LA. 2011. Hemoglobin adducts in workers exposed to 1,6-hexamethylene diisocyanate. Biomarkers. 16(3):261-270.

Flack SL, Fent KW, Trelles Gaines LG, Thomasen JM, Whittaker S, Ball LM, Nylander-French LA. 2010b. Quantitative plasma biomarker analysis in hdi exposure assessment. Ann Occup Hyg. 54(1):41-54.

Gaines LG, Fent KW, Flack SL, Thomasen JM, Ball LM, Richardson DB, Ding K, Whittaker SG, Nylander-French LA. 2010a. Urine 1,6-hexamethylene diamine (hda) levels among workers exposed to 1,6-hexamethylene diisocyanate (hdi). Ann Occup Hyg. 54(6):678-691.

Gaines LG, Fent KW, Flack SL, Thomasen JM, Ball LM, Zhou H, Whittaker SG, Nylander-French LA. 2010b. Effect of creatinine and specific gravity normalization on urinary biomarker 1,6-hexamethylene diamine. J Environ Monit. 12(3):591-599.

Gaines LG, Fent KW, Flack SL, Thomasen JM, Whittaker SG, Nylander-French LA. 2011. Factors affecting variability in the urinary biomarker 1,6-hexamethylene diamine in workers exposed to 1,6-hexamethylene diisocyanate. J Environ Monit. 13(1):119-127.

Gao X, Starmer J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol. 32(4):361-369.

Gentry PR, Hack CE, Haber L, Maier A, Clewell HJ. 2002. An approach for the quantitative consideration of genetic polymorphism data in chemical risk assessment: Examples with warfarin and parathion. Toxicol Sci. 70(1):120-139.

Herrick CA, Xu L, Wisnewski AV, Das J, Redlich CA, Bottomly K. 2002. A novel mouse model of diisocyanate-induced asthma showing allergic-type inflammation in the lung after inhaled antigen challenge. J Allergy Clin Immunol. 109(5):873-878.

Hoffjan S, Nicolae D, Ober C. 2003. Association studies for asthma and atopic diseases: A comprehensive review of the literature. Respir Res. 4:14.

Jiang R, French JE, Stober VP, Kang-Sickel JC, Zou F, Nylander-French LA. 2012. Single-nucleotide polymorphisms associated with skin naphthyl-keratin adduct levels in workers exposed to naphthalene. Environ Health Perspect. 120(6):857-864.

Kim SH, Cho BY, Park CS, Shin ES, Cho EY, Yang EM, Kim CW, Hong CS, Lee JE, Park HS. 2009. Alpha-t-catenin (ctnna3) gene was identified as a risk variant for toluene diisocyanate-induced asthma by genome-wide association analysis. Clin Exp Allergy. 39(2):203-212.

Lange RW, Day BW, Lemus R, Tyurin VA, Kagan VE, Karol MH. 1999. Intracellular s-glutathionyl adducts in murine lung and human bronchoepithelial cells after exposure to diisocyanatotoluene. Chem Res Toxicol. 12(10):931-936.

Liu Q, Wisnewski AV. 2003. Recent developments in diisocyanate asthma. Ann Allergy Asthma Immunol. 90(5 Suppl 2):35-41.

Maestrelli P, Boschetto P, Fabbri LM, Mapp CE. 2009. Mechanisms of occupational asthma. J Allergy Clin Immunol. 123(3):531-542; quiz 543-534.

Mapp CE, Fryer AA, De Marzo N, Pozzato V, Padoan M, Boschetto P, Strange RC, Hemmingsen A, Spiteri MA. 2002. Glutathione s-transferase gstp1 is a susceptibility gene for occupational asthma induced by isocyanates. J Allergy Clin Immunol. 109(5):867-872.

NIOSH. 2015. Hexamethylene diisocyanate. Icsc:0278. In: CDC, editor.

Nylander-French LA, Wu MC, French JE, Boyer JC, Smeester L, Sanders AP, Fry RC. 2014. Dna methylation modifies urine biomarker levels in 1,6-hexamethylene diisocyanate exposed workers: A pilot study. Toxicol Lett. 231(2):217-226.

Ober C, Hoffjan S. 2006. Asthma genetics 2006: The long and winding road to gene discovery. Genes Immun. 7(2):95-100.

OSHA. 2006. Regulations (standards - 29 cfr) part 1910.134. Respiratory protection. Occupational safety and health standards.

Parajuli RP, Goodrich JM, Chou HN, Gruninger SE, Dolinoy DC, Franzblau A, Basu N. 2015. Genetic polymorphisms are associated with hair, blood, and urine mercury levels in the american dental association (ada) study participants. Environ Res. [Epub ahead of print].

Piirilä PL, Nordman H, Keskinen HM, Luukkonen R, Salo SP, Tuomi TO, Tuppurainen M. 2000. Long-term follow-up of hexamethylene diisocyanate-, diphenylmethane diisocyanate-, and toluene diisocyanate-induced asthma. Am J Respir Crit Care Med. 162(2 Pt 1):516-522.

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics. 7:216.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559-575.

Rojas D, Rager JE, Smeester L, Bailey KA, Drobná Z, Rubio-Andrade M, Stýblo M, García-Vargas G, Fry RC. 2015. Prenatal arsenic exposure and the epigenome: Identifying sites of 5-methylcytosine alterations that predict functional changes in gene expression in newborn cord blood and subsequent birth outcomes. Toxicol Sci. 143(1):97-106.

Schulte PA, Whittaker C, Curran CP. 2015. Considerations for using genetic and epigenetic information in occupational health risk assessment and standard setting. J Occup Environ Hyg. 12 Suppl 1:S69-81.

Simino J, Shi G, Bis JC, Chasman DI, Ehret GB, Gu X, Guo X, Hwang SJ, Sijbrands E, Smith AV et al. 2014. Gene-age interactions in blood pressure regulation: A large-scale investigation with the charge, global bpgen, and icbp consortia. Am J Hum Genet. 95(1):24-38.

Thomasen JM, Fent KW, Nylander-French LA. 2011a. Development of a sampling patch to measure dermal exposures to monomeric and polymeric 1,6-hexamethylene diisocyanate: A pilot study. J Occup Environ Hyg. 8(12):709-717.

Thomasen JM, Fent KW, Reeb-Whitaker C, Whittaker SG, Nylander-French LA. 2011b. Field comparison of air sampling methods for monomeric and polymeric 1,6-hexamethylene diisocyanate. J Occup Environ Hyg. 8(3):161-178.

Thomasen JM, Nylander-French LA. 2012. Penetration patterns of monomeric and polymeric 1,6-hexamethylene diisocyanate monomer in human skin. J Environ Monit. 14(3):951-960.

Tijsterman M, Plasterk RH. 2004. Dicers at risc; the mechanism of rnai. Cell. 117(1):1-3.

von Mutius E. 2009. Gene-environment interactions in asthma. J Allergy Clin Immunol. 123(1):3-11; quiz 12-13.

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT et al. 2010. The genemania prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 38(Web Server issue):W214-220.

Weber-Boyvat M, Zhong W, Yan D, Olkkonen VM. 2013. Oxysterol-binding proteins: Functions in cell regulation beyond lipid metabolism. Biochem Pharmacol. 86(1):89-95.
Wisnewski AV, Lemus R, Karol MH, Redlich CA. 1999. Isocyanate-conjugated human lung epithelial cell proteins: A link between exposure and asthma? J Allergy Clin Immunol. 104(2 Pt 1):341-347.
Wisnewski AV, Redlich CA. 2001. Recent developments in diisocyanate asthma. Curr Opin Allergy Clin Immunol. 1(2):169-175.
Wisnewski AV, Stowe MH, Nerlinger A, Opare-Addo P, Decamp D, Kleinsmith CR, Redlich CA. 2012. Biomonitoring hexamethylene diisocyanate (hdi) exposure based on serum levels of hdi-specific igg. Ann Occup Hyg. 56(8):901-910.
Xiao R, Boehnke M. 2011. Quantifying and correcting for the winner's curse in quantitative-trait association studies. Genet Epidemiol. 35(3):133-138.
Yucesoy B, Johnson VJ, Lummus ZL, Kashon ML, Rao M, Bannerman-Thompson H, Frye B, Wang W, Gautrin D, Cartier A et al. 2014. Genetic variants in the major histocompatibility complex class i and class ii genes are associated with diisocyanate-induced asthma. J Occup Environ Med. 56(4):382-387.
Yucesoy B, Kaufman KM, Lummus ZL, Weirauch MT, Zhang G, Cartier A, Boulet LP, Sastre J, Quirce S, Tarlo SM et al. 2015. Genome-wide association study identifies novel loci associated with diisocyanate-induced occupational asthma. Toxicol Sci. 146(1):192-201.

Appendix A: PLINK Codes

Genome-wide Analysis

Raw data from Affymetrix software was output to PLINK format (PED, MAP files) which were then converted to binary BED (and BIM and FAM) files for faster processing in PLINK.

```
plink --file output3_ALL_standard_plink_24Sep2014 --make-bed --out
output3_ALL_standard_plink_24Sep2014
```

```
plink --bfile output3_ALL_standard_plink_24Sep2014 --hwe 0.001 --mind 0.1 --
maf 0.1 --geno 0.1  --autosome --make-bed --out 56_final_10p_12Apr2015
```

The covariate file used is called covar7_15Jan2015.txt. The phenotype file is called GeoMeans_labels_15Dec2014 and includes the geometric mean values of all of the exposure and biomarker variables.

Model 1: c_Smoker,everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm
Model 2: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm
Model 3: everSmoke, DayHDIP_APF_gm,DaySkin_HDI_gm
Model 4: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity
Model 5: everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity
Model 6: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,booth_down
Model 7: everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,booth_down
Model 8: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,cov
Model 9: everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,cov
Model 10: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,cov,booth_down
Model 11: everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,cov,booth_down
Model 12: c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,cov,booth_down
Model 13: everSmoke,DayHDIP_APF_gm,DaySkin_HDI_gm,cov,booth_down
Model 14: DayHDIP_APF_gm,DaySkin_HDI_gm
Model 15: no covariates

The final PLINK association code used included c_smoker, DayHDIP_APF_gm, dayskin_hdi_gm, and Ethnicity as covariates.

```
Plink9 --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt -
-pheno-name UTcr_HDA_gm --pfilter 1e-3 --linear --adjust --covar
covar7_15Jan2015.txt --covar-name
DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,c_Smoker --out utcr_covar_6apr2016
```

```
Plink9 --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --pheno-name Blood_Tot_gm --pfilter 1e-3 --linear --adjust --covar covar7_15Jan2015.txt --covar-name DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,c_Smoker --out bloodtot_newphen_6apr2016

Plink9 --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --pheno-name P_HDA_gm --pfilter 1e-3 --linear --adjust --covar covar7_15Jan2015.txt --covar-name DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,c_Smoker --out plasma_newphen_6apr2016

Plink9 --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --pheno-name Hb_HDA_gm --pfilter 1e-3 --linear --adjust --covar covar7_15Jan2015.txt --covar-name DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity,c_Smoker --out hb_newphen_6apr2016
```

PLINK was used to generate the MDS matrix and the first three columns from the output were used as a proxy for population substructure.

```
Plink --bfile 56_final_10p_12Apr2015 --pheno geomeans_labels_15Dec2014.txt --pheno-name UTcr_HDA_gm --genome --out genome10_21Apr2015

Plink --bfile 56_final_10p_12Apr2015 --pheno geomeans_labels_15Dec2014.txt --pheno-name UTcr_HDA_gm --read-genome genome10_21Apr2015.genome --cluster --mds-plot 4 --out 4mds_21Apr2015
```

However, the MDS vectors did not correlate well with self-reported ethnicity so a binary non-Hispanic Caucasian (coded as 0) or other ethnicity (coded as 1) variable was used instead.

Subsequent GWAS analyses were all in SAS (e.g.., annotation, merging with demographics data, transforming data to run mixed models, running the mixed models). The same files and PLINK codes were used for the genome-wide association for plasma, blood total, and hemoglobin HDA levels, with the --pheno-name changed to reflect the chosen phenotype.

Candidate Gene Analysis

The following command matches SNPs from the uncleaned genotyping dataset to the HG18 annotation and identifies SNPs in our data that are within +/- 20 kb of any annotated gene

```
plink -- output3_ALL_standard_plink_24Sep2014  --make-set glist-hg18.txt --make-set-border 20 --write-set

Plink --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --pheno-name UTcr_HDA_gm --linear --subset SNPlist_2.txt --set plink.set --covar covar7_15Jan2015.txt --covar-name c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity  --adjust --out UTcr_genelist2_13apr2016
```

```
Plink --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --
pheno-name Hb_HDA_gm --linear --subset SNPlist_2.txt --set plink.set --covar
covar7_15Jan2015.txt --covar-name
c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity  --adjust --out
Hb_genelist2_13apr2016

Plink --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --
pheno-name P_HDA_gm --linear --subset SNPlist_2.txt --set plink.set --covar
covar7_15Jan2015.txt --covar-name
c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity  --adjust --out
P_genelist2_13apr2016

Plink --bfile 56_final_10p_12Apr2015  --pheno geomeans_labels_6apr2016.txt --
pheno-name Blood_Tot_gm --linear --subset SNPlist_2.txt --set plink.set --
covar covar7_15Jan2015.txt --covar-name
c_Smoker,DayHDIP_APF_gm,DaySkin_HDI_gm,Ethnicity  --adjust --out
BloodTot_genelist2_13apr2016
```

Appendix B: SAS Codes

Creating Covariate and Phenotype Files For PLINK

```
LIBNAME genetic ('C:\Users\kathie\Documents\Isocyanate data\fall 2015'
'C:\Users\kathie\Documents\Isocyanate data\Genetic data\Phenotypes');


/*read in full study data for n=56, minus one urine measurement for Worker
IID 2913, as genetic.pheno_new*/

proc print data=genetic.pheno_new; run;

DATA eachWorker;
      SET genetic.pheno_new;
      newFID_2 = CATS('FAM', FID, '01');
      DROP FID;
      RENAME newFID_2 = FID;
      eachWorker = CATS(shop, worker);
RUN;

PROC SORT data=eachWorker;
      BY eachWorker visit;
RUN;
proc print data=eachworker; run;

DATA transpose;
      SET eachWorker;
      BY eachWorker;

      LENGTH IIDnew Cell_Datanew DNA_file__new SNP6_file_IDnew $18;

      ARRAY measurements {*}
```

```
      DayAir_HDI DayAirT_HDI DayAirP_HDI DayHDIT_APF DayHDIP_APF
      DayAir_ISO DayAirT_ISO DayAirP_ISO DayISOT_APF
      DayISOP_APF DaySkin_HDI DaySkin_ISO DayPaint_HDI DayPaint_ISO
      Uavg_HDA UT_HDA Uavgcr_HDA UTcr_HDA
      Uavgsg_HDA UTsg_HDA Pconc_HDA P_HDA Pamt_BSA_HDA Pamt_BW_HDA
      Hb_HDA Hb_amt_BSA_HDA Hb_amt_BW_HDA
      Blood_Tot Blood_Tot_Dose;

ARRAY measOne{29};
ARRAY measTwo{29};
ARRAY measThree{29};
ARRAY measFour{29};

RETAIN measOne1-measOne29 measTwo1-measTwo29 measThree1-measThree29
measFour1-measFour29
      IIDnew Cell_Datanew DNA_file__new SNP6_file_IDnew;

IF first.eachWorker THEN DO i=1 to 29;
      measOne{i}=.;
      measTwo{i}=.;
      measThree{i}=.;
      measFour{i}=.;
      IIDnew='';
      Cell_Datanew='';
      DNA_file__new='';
      SNP6_file_IDnew='';
      END;
IF visit=1 THEN DO i=1 to 29;
      measOne{i}=measurements{i};
            IF MISSING(IID)=0 THEN DO;
            IIDnew=IID;
            Cell_Datanew=Cell_Data;
            DNA_file__new=DNA_file__;
            SNP6_file_IDnew=SNP6_file_ID;
            END;
      END;
IF visit=2 THEN DO i=1 to 29;
      measTwo{i}=measurements{i};
            IF MISSING(IID)=0 THEN DO;
            IIDnew=IID;
            Cell_Datanew=Cell_Data;
            DNA_file__new=DNA_file__;
            SNP6_file_IDnew=SNP6_file_ID;
            END;
      END;
IF visit=3 THEN DO i=1 to 29;
      measThree{i}=measurements{i};
            IF MISSING(IID)=0 THEN DO;
            IIDnew=IID;
            Cell_Datanew=Cell_Data;
            DNA_file__new=DNA_file__;
            SNP6_file_IDnew=SNP6_file_ID;
            END;
      END;
IF visit=4 THEN DO i=1 to 29;
      measFour{i}=measurements{i};
            IF MISSING(IID)=0 THEN DO;
```

```
                  IIDnew=IID;
                  Cell_Datanew=Cell_Data;
                  DNA_file__new=DNA_file__;
                  SNP6_file_IDnew=SNP6_file_ID;
                  END;
            END;

      IF last.eachWorker THEN DO;
            OUTPUT;
      END;
RUN;

DATA count;
      SET transpose;
      WHERE MISSING(IIDnew)=0;
      IF MISSING(measOne15)=0 or MISSING(measOne21)=0;
      LENGTH n 3;
      RETAIN n 0;
      n=n+1;

      DROP DayAir_HDI DayAirT_HDI DayAirP_HDI DayHDIT_APF DayHDIP_APF
      DayAir_ISO DayAirT_ISO DayAirP_ISO DayISOT_APF
      DayISOP_APF DaySkin_HDI DaySkin_ISO DayPaint_HDI DayPaint_ISO Uavg_HDA
      UT_HDA Uavgcr_HDA UTcr_HDA
      Uavgsg_HDA UTsg_HDA Pconc_HDA P_HDA Pamt_BSA_HDA Pamt_BW_HDA Hb_HDA
      Hb_amt_BSA_HDA Hb_amt_BW_HDA
      Blood_Tot Blood_Tot_Dose IID Cell_Data DNA_file__ SNP6_File_ID workerid
      shop worker i;
      OUTPUT;
RUN;

DATA genetic.geomeans_logsums_33;
      SET count;
      BY eachWorker;
      RETAIN n eachWorker IIDnew Cell_Datanew DNA_file__new SNP6_file_IDnew;
      LABEL eachWorker='Unique Worker ID' IIDnew='IID' Cell_Datanew='Cell
Data' DNA_file__new='DNA File #'
                  SNP6_file_IDnew='SNP6 File ID';
      ARRAY measOne{29} measOne1-measOne29;
      ARRAY measTwo{29} measTwo1-measTwo29;
      ARRAY measThree{29} measThree1-measThree29;
      ARRAY measFour{29} measFour1-measFour29;
      ARRAY GMmeasurements {*}
            DayAir_HDI_gm DayAirT_HDI_gm DayAirP_HDI_gm DayHDIT_APF_gm
            DayHDIP_APF_gm DayAir_ISO_gm DayAirT_ISO_gm
            DayAirP_ISO_gm DayISOT_APF_gm DayISOP_APF_gm DaySkin_HDI_gm
            DaySkin_ISO_gm DayPaint_HDI_gm DayPaint_ISO_gm
            Uavg_HDA_gm UT_HDA_gm Uavgcr_HDA_gm UTcr_HDA_gm Uavgsg_HDA_gm
            UTsg_HDA_gm Pconc_HDA_gm P_HDA_gm
            Pamt_BSA_HDA_gm Pamt_BW_HDA_gm Hb_HDA_gm Hb_amt_BSA_HDA_gm
            Hb_amt_BW_HDA_gm Blood_Tot_gm Blood_Tot_Dose_gm;
      ARRAY sum_measurements {*}
            DayAir_HDI_m DayAirT_HDI_m DayAirP_HDI_m DayHDIT_APF_m
            DayHDIP_APF_m DayAir_ISO_m DayAirT_ISO_m
            DayAirP_ISO_m DayISOT_APF_m DayISOP_APF_m DaySkin_HDI_m
            DaySkin_ISO_m DayPaint_HDI_m DayPaint_ISO_m
```

```
            Uavg_HDA_m UT_HDA_m Uavgcr_HDA_m UTcr_HDA_m Uavgsg_HDA_m
            UTsg_HDA_m Pconc_HDA_m P_HDA_m
            Pamt_BSA_HDA_m Pamt_BW_HDA_m Hb_HDA_m Hb_amt_BSA_HDA_m
            Hb_amt_BW_HDA_m Blood_Tot_m Blood_Tot_Dose_m;
      IF first.eachWorker THEN DO i=1 TO 29;
            GMmeasurements{i}=0;
            END;
      DO i=1 TO 29;
            GMmeasurements{i} = GEOMEAN(measOne{i}, measTwo{i}, measThree{i},
            measFour{i});
            sum_measurements{i} = sum(log(measOne{i}), log(measTwo{i}),
            log(measThree{i}), log(measFour{i}));
            END;
      DROP measOne1-measOne29 measTwo1-measTwo29 measThree1-measThree29
      measFour1-measFour29 i
      Half_faceRep Half_faceDisp Full_faceCart Air_supply HoodPAPR Cov
      Cov_mat Gloves Glov_tp Glov_thk Hat
      Goggles Booth_tp Gun_HVLP Visit No_Tasks DayTotal_time
      DayPaint_time Clearcoat FID eachworker
      Cell_Datanew DNA_file__new SNP6_file_IDnew n;
      rename iidnew=iid;
      OUTPUT;
RUN;


ods html; PROC PRINT DATA=genetic.geomeans_logsums_33 noobs label;
TITLE '33 participants with genotyping and biomarker data: Geometric Means
and Log-transformed Sums of Exposure and Biomarker Measurements';
RUN; ods html close;

/*Descriptive statistics of geometric mean values of measurements*/
DATA genetic.gm_new_pheno;
      SET means;
      KEEP FID IIDnew Uavg_HDA_gm UT_HDA_gm Uavgcr_HDA_gm UTcr_HDA_gm
      Uavgsg_HDA_gm UTsg_HDA_gm Pconc_HDA_gm P_HDA_gm
      Pamt_BSA_HDA_gm Pamt_BW_HDA_gm Hb_HDA_gm Hb_amt_BSA_HDA_gm
      Hb_amt_BW_HDA_gm Blood_Tot_gm Blood_Tot_Dose_gm;
RUN;

ods csv; PROC PRINT DATA=genetic.gm_new_pheno;
RUN; ods csv close;
/*this is the final dataset for PLINK analysis; includes geometric mean
values of HDA biomarker and HDI exposure measurements*/
```

Annotation

SNPs associated with biomarker levels were read into SAS using the .adjusted files output by PLINK so that the markers were listed in descending order of significance.

```
/*limit genotyping data to 33 workers with complete genotyping and
biomarker/exposure measurements*/
data genetic.genotyping_33_2apr2016;
      set genetic.all_56_genotyping_2apr2016;
```

```
      keep probe_Set_id _0112B _0212B _0613B _0722B _1012B _1211B _1221B
      _1311B _1411B _1713B _1813B _1822B _1913B _1922B _1933B _2013B _2111B
      _2211B _2321B _2411B _2513B _2612B _2713B _2811B _2913B _2922B _3013B
      _3023B _3113B _3312B _3412B _3511B _3612B;
run;

/********** merges genotyping results (significant SNPs in blood and urine)
and annotations downloaded from Affymetrix website *******/
proc sort data=genetic.genotyping_33_2apr2016;
by probe_set_id;
run;

data alleles_urine_blood;
      merge genetic.genotyping_33_2apr2016
      genetic.new_blood_sigsnps_6apr2016(rename=(snp=probe_set_id) in=inlist)
      genetic.new_utcr_sigsnps_30_6apr2016(rename=(snp=probe_set_id)
      in=inlist2) genetic.annotate1 (keep=probe_set_id dbsnp_rs_id);
      by probe_set_id;
      if missing(dbsnp_rs_id)=0;
      if inlist or inlist2;
      if fdr_bh < 0.20;
run;
/********** new_blood_sigsnps_6apr2016, new_utcr_sigsnps_30_6apr2016 are
outputs from PLINK read into SAS datasets ********/

proc sort data=alleles_urine_blood; by unadj;
run;

ods csv; proc print data=alleles_urine_blood;
run; ods csv close;
```

Create Datasets for Linear Regression

Annotated SNPs with significance controlled for FDR<0.20 were output as CSV files and transposed in Excel so that genetic markers could be used as covariates. This file was then read into a SAS database called genetic.trans_newphen_allbiom_7apr2016.

```
proc sort data=genetic.geomeans_logsums_33;
by iid; run;

proc sort data=genetic.covarmds;
by iid; run;


/* Create dataset for multiple linear regression data:
Exposures and biomarkers (both (1) geometric mean values and (2) natural log-
transformed cumulative measurements),genotyping data, annotations */
data complete_lin_model_data;
      merge genetic.trans_newphen_allbiom_7apr2016
      genetic.geomeans_logsums_33 genetic.covarMDS (keep= iid c_smoker
      eth_dum ethnicity);
      by iid;
      count+1;
```

```
      label utcr_HDA_gm='GM log-urine HDA, creatinine adj'
            p_hda_gm = 'GM log-plasma HDA'
            hb_hda_gm = 'GM log-hemoglobin HDA'
            blood_tot_gm = 'GM log-blood total HDA'
            dayhdip_apf_gm = 'GM BZC HDI paint time- and APF-adj'
            dayskin_hdi_gm = 'GM Skin HDI'
            DayHDIP_APF_m = 'Log-cumulative BZC HDI paint time- and APF-adj'
            DaySkin_HDI_m = 'Log-cumulative Skin HDI'
            count = "Worker #";
run;

ods csv ; proc print data=complete_lin_model_data;
run; ods csv close;

title 'Normality of geometric mean of measurements';
proc univariate data=complete_lin_model_data plots normal;
      var UTcr_HDA_gm P_HDA_gm Hb_HDA_gm Blood_Tot_gm
      qqplot;
run;

/******************/
/* Create dataset for multiple linear regression data:
Log-transformed repeated measurements for exposures and biomarkers,genotyping
data, annotations */

data log_exposures_33;
      set genetic.PHENO_NEW_33;
      ARRAY measurements {*}
            DayAir_HDI DayAirT_HDI DayAirP_HDI DayHDIT_APF DayHDIP_APF
DayAir_ISO DayAirT_ISO DayAirP_ISO DayISOT_APF
            DayISOP_APF DaySkin_HDI DaySkin_ISO DayPaint_HDI DayPaint_ISO
Uavg_HDA UT_HDA Uavgcr_HDA UTcr_HDA
            Uavgsg_HDA UTsg_HDA Pconc_HDA P_HDA Pamt_BSA_HDA Pamt_BW_HDA
Hb_HDA Hb_amt_BSA_HDA Hb_amt_BW_HDA
            Blood_Tot Blood_Tot_Dose;
      ARRAY ln_meas {*}
            lnDayAir_HDI lnDayAirT_HDI lnDayAirP_HDI lnDayHDIT_APF
lnDayHDIP_APF lnDayAir_ISO lnDayAirT_ISO lnDayAirP_ISO lnDayISOT_APF
            lnDayISOP_APF lnDaySkin_HDI lnDaySkin_ISO lnDayPaint_HDI
lnDayPaint_ISO lnUavg_HDA lnUT_HDA lnUavgcr_HDA lnUTcr_HDA
            lnUavgsg_HDA lnUTsg_HDA lnPconc_HDA lnP_HDA lnPamt_BSA_HDA
lnPamt_BW_HDA lnHb_HDA lnHb_amt_BSA_HDA lnHb_amt_BW_HDA
            lnBlood_Tot lnBlood_Tot_Dose;
      DO i=1 TO 29;
            ln_meas{i}=LOG(measurements{i});
      END;
      DROP i;
      OUTPUT;

run;

data complete_mix_model_data;
      merge log_exposures_33 genetic.trans_newphen_allbiom_7apr2016
genetic.covarMDS genetic.covarMDS (keep= iid c_smoker eth_dum ethnicity);
      by iid;
      if first.iid then count+1;
run;
```

```
ods html; proc print data=complete_mix_model_data;
run;

title 'Normality of log-transformed measurements';
ods html; proc univariate data=complete_mix_model_data plots normal;
      var lnUTcr_HDA lnP_HDA lnHb_HDA lnBlood_Tot;
      label lnUT_HDA='Log-transformed urine HDA, adjusted for creatinine'
            lnp_hda = 'Log-transformed plasma HDA'
            lnhb_hda = 'Log-transformed hemoglobin HDA'
            lnblood_tot = 'Log-transformed sum of plasma and hemoglobin HDA';
      qqplot;
 run;

title 'Normality of un-transformed measurements';
proc univariate data=complete_mix_model_data plots normal;
      var Uavg_HDA UT_HDA Uavgcr_HDA UTcr_HDA
      Uavgsg_HDA UTsg_HDA Pconc_HDA P_HDA Pamt_BSA_HDA Pamt_BW_HDA Hb_HDA
      Hb_amt_BSA_HDA Hb_amt_BW_HDA Blood_Tot Blood_Tot_Dose;
      qqplot;
run;
```

Linear Models

Macros for Running Linear Regressions

```
%MACRO check_plink(snp= , filen=, marker=);
/*Multiple regression with geometric mean values; run diagnostics on PLINK
analyses*/

proc glm data=&filen plots=(diagnostics(label) residuals);
      class &snp;
      model &marker=&snp c_Smoker DayHDIP_APF_gm DaySkin_HDI_gm eth_dum;
      means &snp;
      output out=out_&snp rstudent=studresid predicted=predicted h=leverage
      cookd=inf predicted=fit residual=resid;
      title "Association of &marker with &SNP";
      title2 'Covariates: Current smoking status, air HDI (paint-time, APF-
      adjusted), skin HDI, ethnicity';
run;

proc sort data=out_&snp;
      by descending inf ; run;

proc print data=out_&snp (obs=5) label;
      var iid inf &marker DayHDIP_APF_m DaySkin_HDI_m;
      label inf='D';
      label utcr_HDA_gm='GM log-urine HDA, creatinine adj'
            p_hda_gm = 'GM log-plasma HDA'
            hb_hda_gm = 'GM log-hemoglobin HDA'
            blood_tot_gm = 'GM log-blood total HDA';
      title "Cook's distance by worker with &SNP"; run;

PROC MEANS data=&filen;
      CLASS &SNP;
```

```
      VAR &marker;
RUN;

PROC SGPLOT DATA=&filen;
      VBOX &marker / CATEGORY=&SNP ;
      title "Distribution of &marker";
RUN;

Proc sgplot data=out_&snp;
scatter y=resid x=fit / label=count;
title "Predicted vs residuals, &snp";
run;

Proc sgplot data=out_&snp;
scatter y=studresid x=fit ;
title "Predicted vs studentized residuals, &snp";
run;

%MEND;

%MACRO lev(snp= , filen=, marker=, pnum=, n=);
/*calculate leverage for points in multiple regression with geometric mean
values (checking PLINK analyses)*/

proc sort data=out_&snp;
by descending leverage ; run;


data out_l;
set out_&snp (obs=10);
p=&pnum; n=&n;
F=((leverage-(1/n)/(p-1)))/((1-leverage)/(n-p));
pvalue=1-probf(F,p-1,n-p);
if pvalue <=0.05/n then BONF="*";
else BONF=" "; *Bonferroni correction;
label Bonf="Signif at 0.05/n?";
run;

ods html; proc print data=out_l uniform label noobs;
var count &marker &snp DayHDIP_APF_gm DaySkin_HDI_gm leverage F pvalue Bonf;
title "Leverage of data points, &marker";
run;
%mend;

%MACRO linreg(snp= , filen= , marker=);
/*multiple linear regression with cumulative measurements for exposure models
with blood biomarkers*/

proc glm data=&filen plots=(diagnostics(label) residuals);
      class &snp;
      model &marker=&snp DayHDIP_APF_m / solution;
      means &snp;
      output out=out_&snp rstudent=studresid predicted=predicted h=leverage
      cookd=inf predicted=fit residual=resid;
      title "Exposure model of &marker with &SNP";
      title2 'Covariates: Cumulative air HDI (paint-time, APF-adjusted),
      cumulative skin HDI';
```

```
run;
%MEND;
```

Multiple Linear Regression Models

```
ods graphics on;

/*DIAGNOSTIC PLOTS FOR TOP 5 SNPS FOR URINE*/
%check_plink(snp=rs169, filen=complete_lin_model_data, marker=utcr_hda_gm)
%check_plink(snp= rs9565949, filen=complete_lin_model_data,
marker=utcr_hda_gm)
%check_plink(snp= rs17472697, filen=complete_lin_model_data,
marker=utcr_hda_gm)
%check_plink(snp= rs12670377, filen=complete_lin_model_data,
marker=utcr_hda_gm)
%check_plink(snp=  rs9921983, filen=complete_lin_model_data,
marker=utcr_hda_gm)

/*diagnostics for 5 new tot blood*/
%check_plink(snp=rs10134376, filen=complete_lin_model_data,
marker=blood_tot_gm)
%check_plink(snp= rs6573958, filen=complete_lin_model_data,
marker=blood_tot_gm)
%check_plink(snp= rs6573948, filen=complete_lin_model_data,
marker=blood_tot_gm)
%check_plink(snp= rs7155763, filen=complete_lin_model_data,
marker=blood_tot_gm)
%check_plink(snp=  rs6939730, filen=complete_lin_model_data,
marker=blood_tot_gm)

/*diagnostics for plasma*/
%check_plink(snp=rs2061660, filen=complete_lin_model_data, marker=p_hda_gm)
%check_plink(snp= rs2061659, filen=complete_lin_model_data, marker=p_hda_gm)
%check_plink(snp= rs1454322, filen=complete_lin_model_data, marker=p_hda_gm)
%check_plink(snp= rs4870000, filen=complete_lin_model_data, marker=p_hda_gm)

/*exposure models for 5 new tot blood*/
%linreg(snp=rs10134376, filen=complete_lin_model_data, marker=blood_tot_m)
%linreg(snp= rs6573958, filen=complete_lin_model_data, marker=blood_tot_m)
%linreg(snp= rs6573948, filen=complete_lin_model_data, marker=blood_tot_m)
%linreg(snp= rs7155763, filen=complete_lin_model_data, marker=blood_tot_m)
%linreg(snp=  rs6939730, filen=complete_lin_model_data, marker=blood_tot_m)

/*exposure models for plasma*/
%linreg(snp=rs2061660, filen=complete_lin_model_data, marker=p_hda_m)
%linreg(snp= rs2061659, filen=complete_lin_model_data, marker=p_hda_m)
%linreg(snp= rs1454322, filen=complete_lin_model_data, marker=p_hda_m)
%linreg(snp= rs4870000, filen=complete_lin_model_data, marker=p_hda_m)


/*leverage*/
%lev(snp=rs169, filen=complete_lin_model_data, marker=utcr_hda_gm, n=30,
pnum=7 )
%lev(snp=rs10134376, filen=complete_lin_model_data, marker=blood_tot_gm,
n=33, pnum=7)
%lev(snp=rs2061660, filen=complete_lin_model_data, marker=p_hda_gm, n=32,
pnum=6)
```

Linear Mixed Effects Models

```
ods rtf; PROC MIXED data=complete_mix_model_data method=reml covtest;
      CLASS IID rs169 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI rs169/ solution ;
      repeated visit /type=cs subject=iid r rcorr;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, top SNP'; run;

PROC MIXED data=complete_mix_model_data method=reml;
      CLASS IID rs9565949 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI rs9565949/ solution ;
      repeated visit /type=cs subject=iid;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, 2nd SNP'; run;

PROC MIXED data=complete_mix_model_data method=reml;
      CLASS IID rs17472697 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI rs17472697 / solution ;
      repeated visit /type=cs subject=iid;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, 3rd SNP'; run;

PROC MIXED data=complete_mix_model_data method=reml;
      CLASS IID rs12670377 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI rs12670377 / solution ;
      repeated visit /type=cs subject=iid;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, 4th SNP'; run;

PROC MIXED data=complete_mix_model_data method=reml;
      CLASS IID rs9921983 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI rs9921983 / solution ;
      repeated visit /type=cs subject=iid;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, 5th SNP'; run;

PROC MIXED data=complete_mix_model_data method=reml;
      CLASS IID rs169 rs9565949 rs17472697 rs12670377 visit;
      MODEL lnUTcr_HDA = lnDayHDIP_APF lnDaySkin_HDI  rs169 rs9565949
      rs17472697 rs12670377/ solution ;
      repeated visit /type=cs subject=iid;
      title "Linear mixed effects model";
      title2 'Dependent variable: log total urine HDA (creatine-adjusted)';
      title3 'Fixed effects: log air HDI (paint time, and APF-adjusted), log
skin total HDI, top 4 SNPs'; run;
```

Eigenanalysis for Collinearity

```
/*correlations and principle components for cumulative exposure data*/
/*step 1: dummy variables*/
%macro dumdum(rs=, al1=, al2=, al3=);
DATA genetic.dumvars_lin_newphen;
      SET genetic.dumvars_lin_newphen;
      RETAIN &rs._new;
      IF &rs EQ "&al1" THEN &rs._new=0;
      ELSE IF &rs EQ "&al2" THEN &rs._new=1;
      ELSE IF &rs EQ "&al3" THEN &rs._new=2;
      ELSE if &rs EQ ' ' THEN &rs._new=.;
      RUN; %mend;

/*plasma*/
%dumdum (rs=rs2061660, al1=GG, al2=TG, al3=TT)
%dumdum (rs=rs2061659, al1=GG, al2=CG, al3=CC)
%dumdum (rs=rs1454322, al1=TT, al2=TC, al3=CC)
%dumdum (rs=rs4870000, al1=GG, al2=TG, al3=TT)

/*blood tot*/
%dumdum (rs=rs10134376, al1=TT, al2=TC, al3=CC)
%dumdum (rs=rs6573958, al1=TT, al2=TC, al3=CC)
%dumdum (rs=rs6573948, al1=AA, al2=GA, al3=AA)
%dumdum (rs=rs7155763, al1=GG, al2=GA, al3=AA)

/*urine*/
%dumdum (rs=rs169, al1=GG, al2=GA, al3=AA)
%dumdum (rs=rs9565949, al1=AA, al2=GA, al3=GG)
%dumdum (rs=rs17472697, al1=GG, al2=AG, al3=AA)
%dumdum (rs=rs12670377, al1=CC, al2=CA, al3=AA)

/*step 2: run analyses*/
proc print data=genetic.dumvars_lin_newphen; run;
proc corr data=genetic.dumvars_lin_newphen nosimple noprob;
      var lnsum_dayairhdi lnsum_dayskinhdi lnsum_dayairthdi; RUN;
proc princomp data=genetic.dumvars_lin_newphen noint ;
      var lnsum_dayairhdi lnsum_dayskinhdi cov booth_tp; run;

/*blood tot*/
proc princomp data=genetic.dumvars_lin_newphen noint ;
      var rs10134376_new rs6573958_new rs6573948_new rs7155763_new; run;
proc corr data=genetic.dumvars_lin_newphen nosimple noprob ;
      var rs10134376_new rs6573958_new rs6573948_new rs7155763_new; run;

/*blood tot*/
proc princomp data=genetic.dumvars_lin_newphen noint ;
      var rs2061660_new rs2061659_new rs1454322_new rs4870000_new; run;
proc corr data=genetic.dumvars_lin_newphen nosimple noprob ;
var rs2061660_new rs2061659_new rs1454322_new rs4870000_new; run;

/*urine*/
proc princomp data=genetic.dumvars_lin_newphen noint ;
      var rs169_new rs9565949_new rs17472697_new rs12670377_new; run;
proc corr data=genetic.dumvars_lin_newphen nosimple noprob ;
      var rs169_new rs9565949_new rs17472697_new rs12670377_new; run;
```